Molecular Simulations of Conformational Transitions in Biomolecules using a Novel Computational Tool

Giuseppe De Marco

Thesis submitted for the degree of Doctor of Philosophy

University of Sussex

March 2011

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to this or another university for the award of any other degree.

Giuseppe De Marco

Signature:

UNIVERSITY OF SUSSEX

Thesis submitted for the degree of Doctor of Philosophy

Molecular Simulations of Conformational Transitions in Biomolecules using a Novel Computational Tool

Giuseppe De Marco

SUMMARY

The function of biological macromolecules is inherently linked to their complex conformational behaviour. As a consequence, the corresponding potential energy landscape encompasses multiple minima. Some of the intermediate structures between the initial and final states can be characterized by experimental techniques. Computer simulations can explore the dynamics of individual states and bring these together to rationalize the overall process. A novel method based on atomistic structure-based potentials in combination with the empirical valence bond theory (EVB-SBP) has been developed and implemented in the Amber package. The method has been successfully applied to explore various biological processes. The first application of the EVB-SBP approach involves the study of base flipping in B-DNA. The use of simple structure-based potentials are shown to reproduce structural ensembles of stable states obtained by using more accurate force field simulations. Umbrella sampling in conjunction with the energy gap reaction coordinate enables the study of alternative molecular pathways efficiently. The main application of the method is the study of the switching mechanism in a short bistable RNA. Molecular pathways, which connect the two stable states, have been elucidated, with particular interest to the characterisation of the transition state ensemble. In addition, NMR experiments have been performed to support the theoretical findings. Finally, a recent study of large-scale conformational transitions in protein kinases shows the general applicability of the method to different biomolecules.

Acknowledgments

The production of this work has been made possible by the support of many whom I am indebted to. I would like to express my enormous gratitude to my supervisor, Dr Peter Varnai for his constant support, encouragement, inspiration and for all his valuable advice.

This work would not have been possible without the generous funding by the University of Sussex. I would also like to acknowledge the generosity of HPC-Europa, and Francois Duchenne bursaries, both of which funded my research travels, opening doors for fruitful collaborations. On that note, I would like to thank Dr Francesco Gervasio for welcoming me into his lab at the CNIO in Madrid. During this period I was warmly welcomed by many, who have made it a rewarding experience. I would particularly like to thank Dr Ludovico Sutto for all the helpful discussions about this work. My thanks also go to Dr Jean-Louis Leroy for welcoming me in his lab in Paris.

The last three years have been a fruitful experience for me at Sussex. Many people, too many to mention, from the Theoretical Chemistry lab and from the Chemistry department, made this experience very enjoyable.

Along the way I have also met some great individuals and very good friends who left a mark on me. A special thanks goes to Rasha for being my constant support over the last two years.

An enormous thanks to all my close friends, whose humour, and much-felt support reached me all the way from Naples.

I cannot end without thanking my family, whose constant love and encouragement sustained me throughout this journey and got me to where I am now. Thank you to my loving parents, Francesco and Maria-Carmela, my fantastic sister, Carmela and my aunt Giovanna.

Contents

Declaration ..... **ii**

Summary ..... **iii**

Acknowledgments ..... **iv**

Contents ..... **v**

List of Abbreviations ..... **viii**

List of Figures ..... **ix**

List of Tables ..... **xii**

List of Pubblications ..... **xiii**

Introduction **1**

Chapter 1. Structure, dynamics and function of biological macromolecules **4**

1.1 Introduction 4

1.2 Structure of biomolecules 5

1.2.1 Protein structure 6

1.2.2 DNA structure 7

1.2.3 RNA structure 9

1.3 Dynamics of biomolecules 10

1.4 Experimental techniques to study biomolecular structure 11

1.4.1 X-ray crystallography 13

1.4.2 NMR spectroscopy 13

1.5 Summary 15

Chapter 2. Simulation techniques **17**

2.1 Introduction 17

2.2 Quantum mechanical potentials 18

2.3 Empirical force fields 20

2.4 Optimisation of the molecular structure 23

2.5 Solvent models 24

2.6 Molecular dynamics simulation methods 25

2.7 Improved efficiency sampling techniques 27
2.8 Simplified models 29
2.9 Free energy calculations 30
2.9.1 Free energy perturbation 30
2.9.2 Umbrella sampling 31
2.9.3 Metadynamics technique 33
2.10 Empirical valence bond theory 34
2.11 Summary 36

Chapter 3. Development of a novel computational technique 37
3.1 Introduction 37
3.2 Potential energy function 38
3.3 Coupled structure-based potentials 43
3.4 Implementation of the EVB-SBP method 44
3.5 A simple test model 51
3.5.1 Testing the method 51
3.5.2 Parameterisation of the potential energy surface 53
3.6 Improvement of the parameterisation process of the structure-based potential 56
3.7 Summary 58

Chapter 4. Base flipping in a B-DNA 59
4.1 Introduction 59
4.2 Parameterisation of the potential based on structural properties 60
4.2.1 Parameterisation of the potential based on energetic properties 65
4.3 Native state dynamics of a B-DNA 69
4.4 Parameterisation of the two-basins potential 71
4.5 Base flipping in B-DNA: insight into the mechanism 72
4.6 Computational efficiency 78
4.6 Summary 79

Chapter 5. Switching mechanism of a bistable RNA 80
5.1 Introduction 80
5.2 Riboswitches 81
5.3 Bistable RNA sequence 82
5.3.1 Previous experimental data 83
5.4 Two-dimensional modelling of bistable RNA 84

5.4.1 Secondary structure prediction....85
5.4.2 Interconversion mechanism at the secondary structure level....86
5.5 Three-dimensional modelling of stable RNA structures....88
5.6 EVB-SBP simulation of bistable RNA....93
5.6.1 Structure-based potential....94
5.6.2 Parameterisation of the EVB-SBP free energy surface....96
5.6.3 Simulation on the parameterised free energy surface....98
5.7 Free energy map and interconversion pathways....102
5.8 Pseudoknot as the transition state structure....109
5.8 Characterisation of the transition state ensemble....112
5.9 NMR spectroscopy....115
5.9.1 Experimental methods....115
5.9.2 Experimental results....116
5.10 Summary....123

Chapter 6. Concluding remarks and future work....124
6.1 EVB-SBP method and its applications....125
6.2 EVB-SBP method: advantages, limitations and future developments....127
6.3 Future work....128
6.3.1 Preliminary results....129

Bibliography....131

List of Abbreviations

COSY	Correlation Spectroscopy
EVB	Empirical valence bond
FF	Force field
FMN	Flavin mononucleotide
FRET	Fluorescence resonance energy transfer
LES	Locally enhanced sampling
LJ	Lennard Jones
MC	Monte Carlo
MD	Molecular dynamics
MPI	Message passing interface
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
nt	nucleotide
PES	Potential energy surface
PMF	Potential of mean force
PME	Particle mesh Ewald
PK	Pseudoknot
REMD	Replica exchange molecular dynamics
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SAM	S-adenosylmethionine
SBP	Structure-based potential
TPP	Thiamine pyrophosphate
TPS	Transition path sampling
UTR	Untranslated region
VDW	van der Waals
WHAM	Weighted histogram analysis method

List of Figures

Figure 1.1 Main components of each nucleotide unit. 8
Figure 3.1 Definition of native and non-native contacts. 40
Figure 3.2 Lennard-Jones potential between two arbitrary atoms in the system, *i* and *j*.. 41
Figure 3.3 Coupling of diabatic states by using the EVB theory to give a unified potential energy surface. 43
Figure 3.4 File *evb_force*.f of the sander code, where umbrella sampling along an energy gap reaction coordinate is implemented. 46
Figure 3.5 Data flow in the modified sander code [243]. 47
Figure 3.6 Modification to the file morsify.f of the sander code to implement different potential form 48
Figure 3.7 Example of a native contact present in both reference structures with different equilibrium values. 50
Figure 3.8 Modification to the subroutine morsify.f for different definitions of equilibrium values in the LJ potential. 50
Figure 3.9 A two-particle model system to test the EVB-SBP method 51
Figure 3.10 Comparison of the results obtained by numerical simulations using the EVB-SBP method and calculated by analytical expressions. 52
Figure 3.11 Time series of the inter-particle distance and the corresponding probability distribution from equilibrium simulations of a two particle model system. 54
Figure 3.12 The free energy profile of a two particle model system. 54
Figure 3.13 An alternative way to parameterise the structure-based potential. 57
Figure 4.1 Schematic representation of base pairing described by native contacts. 62
Figure 4.2. Schematic representation of the heavy atoms within 4.5 Å from a given base. 63
Figure 4.3 Calculated average root mean square deviation (RMSD) from the canonical B-DNA simulated using the structure-based potential with different cut-off and scaling factors. 64
Figure 4.4 Calculated average per residue root mean square fluctuation (RMSF) from the canonical B-DNA simulated using the structure-based potential with different cut-off and scaling factors. 65
Figure 4.5 A comparison between energy differences evaluated by the MM-PBSA and SBP methods. 67
Figure 4.6 Comparison of RMSD and RMSF calculated along the FF and SBP simulations. 69
Figure 4.7 Comparison of helix inter-base pair parameters calculated along the FF and SBP simulations. 70
Figure 4.8 Schematic representation of the two endpoints of the base flipping process. 71
Figure 4.9 The effect of different simulation parameters in the EVB-SBP method to alter the free energy surface of conformational transitions. 72
Figure 4.10 Diabatic state energies and the adiabatic energy underlying the base flipping process are shown along the energy gap reaction coordinate. 73
Figure 4.11 Free energy changes of base flipping in B-DNA calculated using the EVB-SBP method. 74
Figure 4.12 Opening angle of C18 toward the major and the minor grooves. 75

Figure 4.13 Structural snapshots of C18 flipping toward minor and major groove direction.....76
Figure 4.14 The chi angle is plotted along the energy gap reaction coordinate for two different closing trajectories.....76
Figure 4.15 Comparison of closing trajectories ending in anti or syn conformation of C18. 77
Figure 4.16 Free energy changes of base flipping in B-DNA calculated along an RMSD reaction coordinate with different approaches.....78
Figure 5.1 Schematic mechanism of a riboswitch function involved in gene control. 81
Figure 5.2 Two alternative, stable forms of a designed 20nt RNA sequence. 83
Figure 5.3 Refolding pathways calculated using Kinefold [321].....87
Figure 5.4 Schematic description of model building for the A form and B form RNA hairpin loops.....89
Figure 5.5 Root mean square deviation calculated for all atoms forming the helix, the loop and the single strand is shown along the trajectory for the B form and A form.....91
Figure 5.6 Representative tetraloop structures along the molecular dynamics trajectories. 92
Figure 5.7 Time series of distance calculated between the centres of mass of bases G1 and C20 for the B form and A form. 93
Figure 5.8 Time series of the diabatic energy of RNA hairpin forms B and A.....95
Figure 5.9 Comparison of RMSD calculated along FF and SBP simulations for helix B and helix A.....96
Figure 5.10 Energy profile of the RNA conformational change along the energy gap reaction coordinate 99
Figure 5.11 Free energy changes along the transition pathway between the stable RNA structural forms B and A calculated using the EVB-SBP method.....100
Figure 5.12 Probability distribution of the radius of gyration calculated in “macro-windows” generated from independent umbrella sampling simulations.....101
Figure 5.13 The free energy surface sampled by an individual trajectory started in form B (pathway 1) is shown as a function of the RMSD from the A form helix and from the B form helix.....102
Figure 5.14 The distance between the centres of mass of the heavy atoms involved in the native base pairs of form B is plotted as a function of the energy gap reaction coordinate in pathway 1. 103
Figure 5.15 The all atom RMSD calculated with respect to the loops and helices for states B and A as a function of the energy gap reaction coordinate in pathway 1 to show the unfolding of B form and the refolding of A form.....104
Figure 5.16 T The distance between the centres of mass of the heavy atoms involved in the native base pairs of form A is plotted as a function of the energy gap reaction coordinate in pathway 1. 105
Figure 5.17 The free energy surface sampled by an individual trajectory started in form B (pathway 2) is shown as a function of the RMSD from the A form helix and from the B form helix.....106
Figure 5.18 Structural comparison of the transition state region of pathway 1 and pathway 2. 107
Figure 5.19 The free energy surface sampled by an individual trajectory started in form B (pathway 3) is shown as a function of the RMSD from the A form helix and from the B-form helix. 108
Figure 5.20 Structural comparisons of the transition state region in pathway 3. 108
Figure 5.21 Schematic representations of pseudoknot structures. 109
Figure 5.22 The three-dimensional model of pseudoknot.....110

Figure 5.23 Free energy changes of the interconversion of A and B form RNA via the pseudoknot transition structure. 111

Figure 5.24 P_{fold} values for each structure are shown as a function of RMSD calculated with respect to A and B forms using helix residues and loop residues. 114

Figure 5.25 Gel filtration chromatogram and 1D H-NMR spectrum of the 20nt RNA. 117

Figure 5.26 Imino proton exchange time measurements. 118

Figure 5.27 Arrhenius plot of experimental rate constants [305] for A→B and B→A transitions. 121

Figure 5.28 Imino proton region of H-NMR spectra recorded at various delay times from snap cooling at 288K. 121

Figure 6.1 Comparison of root mean square deviation and root mean square fluctuations calculated along the FF and SBP simulations of a protein system. 130

List of Tables

Table 4.1 Number of native contacts at different values of cut-off for Watson-Crick base pairs 61

Table 4.2 Non-bonded energy components calculated using FF and SBP for six selected pairs of structures. 68

Table 4.3 Average helical parameters and standard deviations of a B-DNA double helix calculated from 50 ns simulated trajectories using force field (FF) and structure-based potential (SBP) 70

Table 5.1 The free energy contributions to secondary structures in the B and A forms calculated by *mfold* [317] 85

Table 5.2 Thermodynamic data for B and A forms calculated using the *mfold* server [317] 86

Table 5.3 Activation parameters for the interconversion of the bistable RNA calculated from experimental data. 97

Table 5.4 Imino proton exchange times. 119

List of Pubblications

De Marco, G. and P. V arnai, Molecular simulation of conformational transitions in biomolecules using a combination of structure-based potential and empirical valence bond theory. Physical Chemistry Chemical Physics, 2009. 11(45): 10694-10700.

Introduction

The last decades have revealed a new fascinating world where small "workers" (biomolecules) are in perpetual activity to ensure the robust cycle of life. As small units of a near perfect machinery, they are connected in a complex network where every component is essential and has a vital role. Understanding the intricate mechanisms underlying this network is of vital importance and represents one of the major challenges of life sciences.

A key example of the strict relation between biomolecules emerged after the outstanding discovery of the DNA [1, 2]. A beautiful picture was revealed where everything is connected in a chain of events starting from the DNA to RNA to proteins: the genetic information is encoded in the DNA and transferred by RNA messenger to be decoded in order to assemble amino acid units into powerful machineries essential for life (proteins).

Notably there are several exciting features, which evidence the complexity and beauty of biological machineries. Biomolecules are extremely cooperative as demonstrated by a variety of key interactions, e.g. protein-protein, protein-nucleic acid and nucleic acid-nucleic acid. The formation of large molecular complexes is a clear example of the importance of cooperation in biomolecules: the proteasome [3, 4] is a large multi-protein complex involved in the degradation of non-functioning proteins; the nucleosome [5], essential for the DNA organization in the cell nucleus, formed by proteins (histone) and DNA; the spliceosome [6], a complex of RNA and protein necessary to eliminate non-coding DNA regions (introns). Biomolecules are very responsive elements, sensitive to external factors, such as temperature, pH or small metabolite concentration. Furthermore, biomolecules are not rigid bodies but highly flexible molecules. In particular, conformational changes often represent an essential prerequisite in order for biomolecules to exert their functions. These and many other features paint a complex picture, where many biological questions are still unanswered.

Understanding the physics and chemistry underlying the function of biological machineries is of fundamental importance. One of the main goals is to elucidate the relationship between structure, dynamics and function. With this intent, both experimental and theoretical approaches have been intensively applied in the last

decades. Experimental techniques such as NMR or X-ray have been efficiently used to study the structure of biomolecules [7, 8]. From the structure of a molecule, predictions on the possible mechanism of actions and functions can be hypothesized. For example, structures of the same molecule in different conformations can give clues to the possible transition mechanism, or enzyme function can be suggested from the structure of a protein-ligand complex. Nevertheless, the study of the dynamic behaviour should be explicitly addressed to gain insights into the role and the functions of these biomolecules, though it remains challenging for experiments to provide an atomic detail of these processes.

Theoretical approaches represent a powerful tool that has been increasingly used in the last decades. Computer simulations are now routinely used, and often complement well experimental studies. Applications cover a wide range of areas including for example structural biology, biochemistry, enzymology, biophysics, medicinal and pharmaceutical chemistry. Theoretical models are used for structure determination and refinement in conjunctions with X-ray [9] and NMR [10] studies. Starting from biomolecular structures, there are several approaches that can be applied to investigate their properties. A common approach is molecular dynamics (MD). MD simulations can be applied to biological systems to study both thermodynamic and kinetic properties. In detail, MD simulations provide a microscopic description of a system in terms of atomic positions and velocities. The use of statistical mechanics then provides a way to connect microscopic properties to macroscopic properties of a system, such as free energy changes.

One of the main limits of standard approaches lies in limitations of simulated time scales. Relevant biological processes occur on the order of milliseconds to seconds or higher, while time scales accessible to classical MD approaches are orders of magnitude lower. This issue mainly arises from the complexity of these calculations and the available computational power. Notably, in the last decade, considerable efforts have been directed in this area following two main lines of research: development of fast supercomputers and of efficient computational techniques.

The goal of the present thesis is to develop an efficient computational tool to study long time-scale conformational transitions in biomolecules. Chapter 1 provides an overview of the main principles of structure, dynamics and function of biomolecules and briefly introduces some experimental techniques which can be applied to study those. Theoretical models, which provide the foundation of the present work, are

described in Chapter 2. The new computational tool is introduced in Chapter 3: theory, implementation and application to a model system are presented. The first application to a relevant biological problem, base flipping in B-DNA, is presented in Chapter 4. Finally, an extensive study of the switching mechanism of a bistable RNA sequence is presented in Chapter 5. This work combines both theory and NMR experiment to elucidate the RNA switching mechanism. Concluding remarks, recent work on a protein kinase and future applications are described in Chapter 6.

Chapter 1

Structure, dynamics and function of biological macromolecules

1.1 Introduction

Biological macromolecules include a large range of molecules, essential for life, such as nucleic acids (DNA, RNA) and proteins. Despite the chemical differences between these molecules, they are functionally interconnected in a complex biological network. The DNA is the repository of genetic information in all cells. It is now well known that the genetic information is transcribed to RNA (transcription) and subsequently translated from RNA to protein (translation). Nucleic acids and proteins can be considered as the structural and functional units of every organism. They play a key role in several biological processes, making it vital to understand their functions. Numerous examples can be given to show the enormous diversity of their specific biological roles [11, 12].

Nucleic acids carry the information from one generation to the next and are necessary to ensure the normal development and function of every organism. The role of reaction catalyst has been addressed initially to proteins and then to RNA (ribozymes). However, recent evidence has shown that even DNA has a potential enzymatic activity, although it has not been found in nature. Catalytic DNA, also known as DNA-zymes, have been discovered in the last decades [13, 14] opening a new exciting field. In addition several studies have demonstrated their successful application in medicinal chemistry, biotechnology and related fields [15-17]. RNA is a unique molecule able to carry both genetic information and catalytic function. It plays a key role in both transfer and processing of genetic information [18]. RNA molecules are involved in several biological mechanisms such as protein synthesis (ribosome) or

catalysis (ribozymes). Furthermore recent theories have suggested a possible involvement of RNA as precursor of life on earth ("RNA world" hypothesis) [19].

Proteins are involved in the transport of several small molecules and ions (e.g., hemoglobin transports oxygen, transferrin carries iron). Antibodies are highly specific proteins, which recognize, combine and neutralize foreign material such as viruses and bacteria. Proteins also regulate cellular growth and differentiation. According to Anfinsen's principle, the sequence encodes the necessary information for folding a protein toward a unique and stable three-dimensional structure [20]. Interestingly, the fold of a protein is more conserved through evolution than its amino acid sequence [21, 22].

A common feature of all biological macromolecules is the strict relation between sequence, structure and function. Since the three-dimensional structure is instrumental in understanding the function, experimental techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR), have been used extensively to determine the time-averaged structures of biomolecules. However, not all biomolecules exhibit well-defined structures. Intrinsically unstructured proteins, for example, are thought to be involved in signal transduction and DNA promoter binding [23-26]. Certain nucleic acid sequences fold into different functional structures [27, 28]. The dynamic nature of macromolecules has recently attracted much attention to bridge the gap between structure and function [29-36]. Nevertheless it still remains challenging to describe the dynamics of these molecules at the atomic level both experimentally and theoretically.

1.2 Structure of biomolecules

Before considering the function of biomolecules, it is necessary to review some of the main features that characterize their three-dimensional organization. There are four levels of hierarchy in biomolecular architecture: primary, secondary, tertiary and quaternary structure [37]. Biomolecules are polymers constituted of different monomeric building blocks, amino acids for proteins and nucleotides for nucleic acids. The sequence of monomeric units is defined as the "primary structure". As introduced above, a particular sequence ultimately determines the three-dimensional structure of proteins ("Anfinsen's dogma"). One of the main goals in structural biology is to predict the structure of a molecule from its primary structure. Recurrent segments of the biopolymer, such as helices and sheets, are called the "secondary structure". The local

spatial arrangement of these segments however does not describe specific atomic positions in the three-dimensional space. Rather, they are defined in terms of hydrogen bonds formed in the biopolymer. The next level in the hierarchy is the packing of the secondary structural elements into one or several independent structural units. This three-dimensional structure defined by the atomic coordinates is called the "tertiary structure". Finally, the "quaternary structure" is related to the higher-level organisation of several subunits or domains, for example, in proteins that contain several polypeptide chains. The principles of protein and nucleic acid structures are outlined in the following.

1.2.1 Protein structure

Proteins are biopolymers built from 20 different amino acids [38]. An amino acid consists of a central carbon atom (C_α) to which are attached a hydrogen atom, an amino group (NH_2), a carboxyl group (COOH), and a side chain (R). There are twenty different side chains specified by the genetic code, which vary in size, charge, hydrophobicity, and chemical reactivity. They can be divided into two main classes: side chains soluble in water, also named hydrophilic, and those that are less soluble in water, termed hydrophobic. The amino acids are joined during protein synthesis by the formation of a peptide bond between the carboxyl group of one amino acid and the amino group of another amino acid. The peptide bond is not free to rotate because of its partial double-bond character. As a consequence, the peptide units are rigid and planar and the main degree of freedom in a polypeptide chain comes from the rotation around the bonds N-C_α (phi) and C_α-C (psi). The rigidity of the peptide unit enables proteins to have a well-defined three-dimensional structure. On the other hand the rotation around the phi and psi angles allows proteins to fold into many different topologies.

The protein main chain (or backbone) is highly polar and hydrophilic due to the presence of one hydrogen bond donor (NH) and one hydrogen bond acceptor (CO) in each peptide unit. However, formation of different secondary structure elements such as α-helix and β-sheet [39, 40], stabilised by hydrogen bonds between NH and CO, contributes to reduce the backbone polarity. The polypeptide chain is tightly coiled in an α-helix, while it is stretched in a β-sheet. The α-helix is stabilized by hydrogen bonds between the NH group of residue *n* and the CO group of residue *n+4* of the main

chain. The β-sheet instead is stabilized by hydrogen bonds between NH and CO groups of different polypeptide strands. Protein structures involve several combinations of α-helices and β-strands to form structural units, called super-secondary structures or motifs [41, 42]. A simple motif can be composed of two helices (helix-turn-helix) or two β-strands (β-hairpin). Another common motif includes two parallel β-strands connected by an α-helix. Structural elements, which form tertiary structures are called domains and are composed by a combination of several motifs. Domain structures are classified in three main groups [43]: α structures, composed exclusively of α-helices; β structures formed only by β−sheets; and α/β structures with alternating α-helices and β-sheets. Usually α-domain structures are composed of a bundle of α-helix packed together to form a hydrophobic core. A common α-structure is the globin fold, present in proteins such as myoglobin or hemoglobin, where eight α-helices pack together to form a pocket where the heme group is bound. One of the most common α/β domain structures is the α/β−barrel (or defined TIM barrel), usually formed of eight β-strands surrounded by an equal number of α-helices to create a common active site for many enzymes. The knowledge of the different structural domains in proteins is very important toward a better understanding of protein evolution and function.

1.2.2 DNA structure

Deoxyribonucleic acid (DNA) is a polymer built from monomeric units called nucleotides (nt). The nucleotide unit is formed of a nitrogenous base, a sugar and a phosphate group. A glycosidic bond is formed between the sugar and the base, while the phosphate groups connect the 3’ carbon of one nucleotide and the 5’ carbon of the other. There are four types of nucleotides, which differ in their nitrogenous base: adenine (A), guanine (G), cytosine (C) and thymine (T) (Fig. 1.1). Bases are planar aromatic rings. Adenine and guanine are purine derivatives containing a pyrimidine ring fused to an imidazole ring, while cytosine and thymidine contain one single pyrimidine ring.

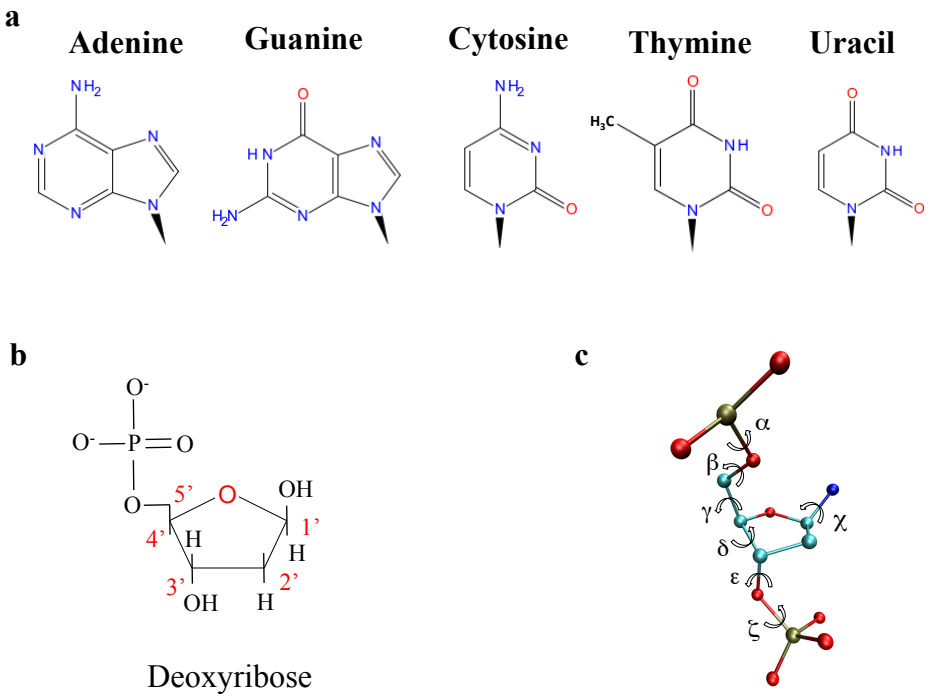

Figure 1.1 Main components of each nucleotide unit: (a) purine and pyrimidine nitrogenous bases; (b) the deoxyribose sugar and the phosphate group; (c) backbone torsional angles.

The DNA double helical form, proposed by Watson and Crick in 1953 [1, 2] is mainly stabilised by hydrogen bonds between nitrogenous bases: A/T pairs form two hydrogen bonds; C/G pair form three hydrogen bonds. However, there are several non-canonical ways to form hydrogen bonds between base pairs [44, 45]. Several DNA structures have been solved using X-ray crystallography and NMR spectroscopy. Right-handed antiparallel double helices [2] are the most common in nature. The backbone consists of two polynucleotide chains coiled along a common axis with opposite directions. Bases are on the inside of the helix with their hydrophobic faces shielded from the solvent by stacking interactions, while the charged phosphate groups are on the outside, fully exposed to the aqueous solvent. Right-handed helices are usually divided into two forms: A-DNA and B-DNA [46]. The form of DNA that is the most stable under physiological condition is the B-DNA [47]. The main factors which stabilize a canonical B-DNA are: 1) enthalpic contribution coming from hydrogen bond formations and stacking between Watson-Crick base pairs; 2) optimal nucleotide conformations with minimal steric clash in the sugar phosphate backbone structure and in the attachment of the base to the backbone. A non-canonical DNA helix is the left-

handed double helix, first discovered in early 1980 [48]. This form, known as Z-DNA, has remarkable differences with respect to the B- and A- DNA and is thermodynamically less stable.

It is worth noting that besides helices there are other possible conformations accessible for DNA. G-quadruplexes are four-stranded structures, formed by nucleic acid sequences rich in guanine [49-52]. These structures are usually present at the end of eukaryotic chromosomes (telomers). G-quadruplexes are essential for chromosome replication and stability and are also involved in a variety of other biological and biochemical processes [53, 54]. Similar to G-quadruplexes are the i-motifs [55] where two opposite parallel strands composed of hemi-protonated cytosines create a first helix. The other two parallel strands then run in opposite directions with their cytosine pairs intercalated between pairs of the first two strands [55]. Another particular conformation is a four-way branched nucleic acid, also known as Holliday junction [56, 57]. These structures occur as intermediates during the genetic recombination of chromosomes. The enormous diversity of DNA structures is also confirmed by the constantly increasing number of new structures deposited in the Protein Data Bank and in the Nucleic Acid Data Bank [58, 59].

1.2.3 RNA structure

Ribonucleic acid (RNA), similar to DNA, is a polymeric chain composed of monomer units (nucleotide), each containing a sugar, a base and a phosphate group. An important structural feature that distinguishes RNA from DNA is the presence of a hydroxyl group at the 2' position of the ribose sugar. In addition, the four major bases in RNA are adenine (A), guanine (G), cytosine (C) and uracil (U). The difference between thymine and uracile is due to the presence of a methyl group on the C5 of the base (Fig 1.1). In contrast to DNA, in which the dominant form includes two strands to form a double helix, RNA is usually a single-stranded polynucleotide chain, able to form double-stranded segments, stabilised by complementary hydrogen bonds between A/U and G/C. Beyond the canonical Watson-Crick base pairs, there are often non-Watson-Crick base pairs [44] or even unpaired bases in RNA. The most common helical form in RNA is the A form. In contrast to the B-form usually observed in DNA, the A-form is shorter and wider, the major groove is deep and narrow and the minor groove is wide and shallow. Notably, RNA has the ability to form complex three-dimensional folds as seen,

for example, in tRNA or ribozymes [60-62]. The structure of RNA can be described as a combination of several independent structural elements, such as hairpin-loop, junction, bulge or internal loop motifs [60-62]. These motifs, mediating tertiary interactions, are stabilized by forming hydrogen bonds between the backbone and bases, or forming base pair triplets. The presence of a wide variety of structural motifs enables the polynucleotide chain to fold into distinct three-dimensional structures.

1.3 Dynamics of biomolecules

Function of biomolecules is strictly dependent on their conformation. The correct three-dimensional structure represents the fundamental prerequisite which allows biomolecules to function correctly. However, understanding the function from single structures is not straightforward. Biological molecules are not static and rigid systems but highly dynamic. Traditional experimental techniques provide a model structure that represents the time-average properties of many molecules. Although these structures can offer several clues about possible function, they do not offer detailed insight into the actual mechanism of action. By connecting instantaneous structures along a conformational transition pathway over the relevant timescales, it becomes possible to characterize the transition state ensemble, essential for understanding mechanistic details of the process.

Biomolecules “never rest” [63] and are in continuous motion between different states. These motions span a wide range of time and spatial scales, from atomic position fluctuations to complex movements, which involve entire domains or subunits. Atomic vibrations are on the subpicosecond time scale, backbone and side-chain/base fluctuations happen on the picosecond to nanosecond time scale, conformational rearrangements on the order of milliseconds, breathing modes on the order of seconds and folding/unfolding on the order of seconds or longer time scales [64]. Any of these motions may be functionally significant: motions of backbone and side-chain atoms may be required for molecular recognition, loop motions may be necessary to expel water or for repositioning catalytic residues [65, 66]. From a different perspective, another general and useful classification [67] distinguishes between two types of motions: relaxation processes and equilibrium fluctuations [67]. Relaxation processes drive transition of a system from a non-equilibrium state toward the equilibrium state. On the other hand, equilibrium fluctuations describe conformational transitions between

different structural states of a molecule. These motions may be necessary to understand the function of a particular biomolecule. It is worth noting that these motions are not random but are strictly dependent on intra- and intermolecular interactions, hence they are "structure-encoded" [68]. To organize the enormous variety of different motions, a database of observed motions in proteins and nucleic acids has been generated [69].

Structural motions are an essential prerequisite for several protein functions, such as catalysis, transport of methabolities, molecular recognitions and many others [70-73]. Particular attention has been paid to the role of conformational dynamics during the chemical reaction in enzymes [71, 74-78].

Local motions in DNA coupled to protein motions intervene in protein recognition mechanisms [79-81]. Following DNA base flipping processes, functional groups, normally buried in the Watson-Crick base pair, are exposed to the solvent becoming accessible to enzyme activity [82-85]. DNA motions (bending, twisting) are essential to regulate the chromosome formation upon DNA interactions with histones, or how it supercoils during replication [86-88]. In comparison, RNA is a flexible polynucleotide chain, which often undergoes complex conformational changes. The many functions of RNA often require change in conformation in response to biological signals such as binding to ions, metabolites, proteins or other nucleic acids [36, 89, 90]. Despite their biological importance, these motions are difficult to identify and characterize quantitatively at the atomistic level. Both experiments and theoretical approaches often fail in describing these motions, which span different length and time scales. In the following, an overview of some experimental techniques is given, which can be applied to study the structure and the dynamics of biomolecules.

1.4 Experimental techniques to study biomolecular structure

Many experimental techniques have been widely applied in the last decades to determine various aspects of structure and dynamics of nucleic acids and proteins. X-ray crystallography was the first technique applied to study the structure of biomolecules at the atomic level [91-93]. At present, more than 68,000 structures of protein, nucleic acid and other biomolecules have been deposited in the protein data bank [58] since the first protein structure was deposited in 1976 [94]. It is interesting to note that almost 60,000 structures (~88%) have been determined using X-Ray.

Additionally, the development of small angle X-ray scattering (SAXS) made it possible to study not only the structure but also the dynamics of biomolecules, in particular, the global changes in the size and shape of biomolecules in a time-resolved manner [95]. SAXS can also be applied to study the kinetics of folding of biomolecules or to obtain low-resolution three-dimensional density maps for either protein or nucleic acid complexes. Following the first protein structures determined by Kendrew et al. [96] and Perutz et al. [92] using X-ray crystallography, Wutrich used NMR spectroscopy to determine the structure of the bovine pancreatic trypsin inhibitor (BPTI) [7]. NMR has proven to be a powerful experimental technique for studying both structure and dynamics. Although X-ray crystallography remains the most widely used approach to determine the structure of large molecular complexes [5, 97-99] NMR spectroscopy is more efficient to determine the structures of small, flexible molecules, which are often difficult to crystallize.

Various other techniques are used to complement X-ray crystallography and NMR spectroscopy to study biomolecular structure and dynamics. Fluorescence spectroscopy in particular, has been widely applied to study the dynamics of proteins and nucleic acids [100, 101]. Fluorescence resonance energy transfer (FRET), measures the efficiency of non-radiative energy transfer between a donor and an acceptor fluorophore. An estimate of the inter-dye distance is derived from the ratio of acceptor intensity to total emission intensity [102]. FRET is able to monitor single molecules (smFRET) [103], revealing dynamic information on the order of millisecond to minutes and is very powerful in describing population distribution of inter-dye distances. One important feature of FRET measurements is that it can be carried out on single molecules [103] not only in vitro but also in a true cellular context [104, 105].

Hydrogen-deuterium exchange mass spectrometry (HX MS) [106-108] has been successfully used to study protein dynamics, specifically in large supramolecular complexes. Finally, more specialized techniques such as Mössbauer spectroscopy [109], Raman spectroscopy [110], and 2D infrared spectroscopy [111] can also provide new insights into biomolecular dynamics. Mössbauer spectroscopy can be applied to study in detail structural, dynamical and electronic properties of iron centres in biomolecules [112, 113]. Raman spectroscopy can be applied to unravel local structure, global conformation and the complex dynamics of nucleic acids [114] and proteins [115]. 2D infrared spectroscopy has been proven very useful due to its ability to provide structural information at high time resolution, and has been used in a variety of applications such

as rapid structural fluctuations, vibrational dynamics and solute-solvent interactions [116]. In the following, a brief description of two of the main techniques is given, which have been extensively applied to study the structure and dynamics of biomolecules.

1.4.1 X-ray crystallography

In order to solve the structure of a protein or nucleic acid by using X-ray diffraction, the first step involves the laborious task of crystallizing the molecule. In the crystal lattice the molecules are arranged in an orderly, repeating pattern in the three dimensions of space. X-ray radiation has a wavelength ranging between 0.1 Å and 100 Å, which enable the determination of the molecular structure at the atomic resolution. The basic concept is that when a wave interacts with a crystal, it will be diffracted provided the repeating lattice distances in the crystal are comparable to the wavelength (Bragg diffraction) [117]. The specific diffraction pattern is dependent on the atomic organization in the crystal. The spatial distribution of electronic density peaks in the crystal allows for the determination of the average positions of the atoms in molecules. The atomic displacement is calculated from the shape of such peaks [118]. In particular the width of the electron density peak is described using the Debye-Waller factor [119] or B-factor, which indicates the mean atomic displacement dependent on thermal fluctuations or vibrational motions within the molecule. Significant progress has been made in data collection and analysis. However, one of the main limitations of X-ray crystallography remains the production of crystals that diffract to a high enough resolution. When crystallization is not possible, then alternative methods of structure determination must be sought, such as NMR spectroscopy that allows determination of the structures of biomolecules in solution.

1.4.2 NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy, together with X-ray crystallography, is one of the most frequently used biophysical methods to provide high-resolution structures of biological molecules. In addition, NMR spectroscopy is useful not only for structure determination but also to study the dynamic behaviour of biomolecules in solution. As a result, temperature, pH and salt concentration can be varied to mimic physiological or even denaturing conditions.

The basic concept of NMR is to exploit the magnetic properties of certain atomic nuclei (i.e., ^{1}H, ^{13}C, ^{15}N, ^{31}P), which have magnetic moment or spin. When a molecule is introduced into a magnetic field, the equilibrium population of a given nuclear spin changes and creates a net magnetization of the sample. Resonance absorption occurs when the frequency of the electromagnetic field corresponds to the energy gap between the nuclear spin levels (Larmor frequency). This is different for different types of nuclei. In practice, a polarized radiofrequency pulse is applied that rotates the magnetization vector of all nuclear spins into a plane perpendicular to the external magnetic field. By measuring the free induction decay signal of the nuclear spins in this plane, Fourier transformation then provides the NMR spectrum of the sample. The resonance frequency is also dependent on the chemical environment around the nuclei and gives rise to chemical shifts with respect to a reference standard. In theory it is possible to obtain a unique chemical shift for each nucleus in the molecule. In practice one of the limitations of classical one-dimensional NMR, especially for complex molecules, is that often the differences in chemical shifts are smaller than the spectral resolution. Particularly useful solutions are the two-dimensional NMR techniques such as the correlation spectroscopy (COSY) [120] and the nuclear overhauser effect spectroscopy (NOESY) [121]. The main idea underlying these approaches is that nuclei are not isolated and their spin can interact between themselves. The COSY experiment is based on the scalar coupling, which derives from the polarisation of the electrons in the orbitals joining the two nuclei. The scalar coupling value decays to zero for nuclei separated by more than 4 bonds. As a result only spin systems, i.e., nuclei separated by no more than three chemical bonds can be observed, using correlation spectroscopy. However, nuclei can also interact through space by sensing the magnetic field created by other nuclei. This effect, also know as nuclear overhauser effect (NOE), depends on the intensity of the interacting magnetic fields, and on the inverse of the sixth power of the distance separating them. The NOESY experiment is able to identify nuclei, which are close in space but not necessarily in sequence, thus overcoming the limitation of the COSY method. Both approaches provide useful information in order to determine the three dimensional structure of the molecule. It is not trivial however, to assign the observed peaks in the spectra to specific residues in the sequence. This can be achieved by using the resonance sequence assignment technique introduced by Wutrich et al. [122].

In addition to structure determination, NMR applications may include the study of dynamic features of the molecules, as well as the kinetic and thermodynamic characterization of the process [123, 124]. The time scales available to NMR techniques covers almost all of the relevant dynamic motions in proteins and nucleic acids. Motions in the order of the picoseconds to nanoseconds are usually studied by measuring relaxation rates such as the longitudinal relaxation rate, R_1, the transverse relaxation rate R_2, and the steady state heteronuclear NOE [125]. In particular, relaxation rates of atomic nuclei are related to dipole-dipole interactions, hence motion of interspin bond vectors can be extracted from relaxation times. This relation can also be expressed in more quantitative terms by introducing the so called spectral density function *J(w)*, which are frequency functions characteristic of the interspin bond vectors motions. In this context the Lipari-Szabo [126] "model free" analysis, is the most common approach for the interpretation of ps-ns dynamics. Beyond nanoseconds, processes in the order of microsecond to millisecond can be studied by measuring the residual dipolar couplings in the sample [127]. Motions beyond seconds can be characterized by hydrogen/deuterium exchange experiments [128, 129]. Finally, the recent development of newer transverse relaxation optimized spectroscopy (TROSY) techniques, have extended the original NMR size limit reaching up to 1000 kDa [130-132].

1.5 Summary

The great importance of biological macromolecules stems from their essential roles in almost every biological process. It is crucial to understand their functions, not only to elucidate the complex mechanisms of these biological machineries, but also because several debilitating diseases arise when these mechanisms are altered. The last century has witnessed numerous discoveries that have profoundly changed our views of life at the cellular and atomic level. In particular, enormous progress has been made in the field of structural biology. The determination of the atomic resolution structures of proteins and nucleic acids, first by X-ray crystallography and successively by NMR spectroscopy, showed the striking shapes these molecules can display. For the first time it was possible to observe the basic elements of life at the atomic level. Beyond the beautiful objects, however, these biomolecules soon revealed the complexity and dynamics of the information flow in biological networks.

Several experimental techniques can be applied to study of structure and dynamics of proteins and nucleic acids. Yet, no universal technique is available at present and thus different methodologies have to be applied to decipher a complex biological problem. Computer simulations at the atomic level can complement effectively the wealth of experimental data available for better characterizing and understanding the working mechanism of functional motions in biomolecules.

In one of his famous lectures on Physics [133], Feynman described the progress in biology in the following way: “all things are made of atoms and that everything that things can do can be understood in terms of jigglings and wigglings of atoms”. The present thesis endeavours to do just that, using a novel computational approach applied to various biological questions.

Chapter 2

Simulation techniques

2.1 Introduction

Computer simulations techniques can be applied to study structure and dynamics of biological macromolecules simultaneously in order to understand their function. The present chapter aims to introduce principal concepts of computer simulations by focusing on methods relevant to the present thesis. Several books that provide broader and deeper insight into the subject are accessible in the literature [64, 134, 135]. Classical molecular dynamics (MD) methods are widely used to describe the changes of the coordinates of particles in a molecular system in time and provides the cornerstone of the present thesis. This approach was first used in the theoretical physics community in the early 1950s. It has been almost 60 years since Alder and Wainwright performed the first MD simulation using the so-called hard-sphere model [136-138]. During the 1970s, as computers became more widespread, MD simulations were developed for more complex systems, culminating in 1977 with the first simulation of a protein, the bovine pancreatic tripsin inhibitor (BPTI) [139] and subsequently with the first simulation of a DNA duplex [140]. These results were crucial to show that biomolecules are dynamical systems and not just rigid structures. Since the very first simulated molecular trajectories, MD approaches have become widely used in many fields of science including structural biology of macromolecules, biophysics, enzymology, pharmaceutical chemistry, and material science. The popularity of the method is explained in that it can serve as a computational microscopy to study the structural, kinetic and thermodynamic properties of the system at an arbitrary resolution of space and time. The analysis of the large body of resulting data by statistical mechanics may reveal fundamental aspects of the biomolecular structure and dynamics, and hence recognition and function.

In this chapter, the description of some basic concepts of computer simulations will start with the high-accuracy models (quantum chemistry) and progress toward more simplified representations that allow practical simulations of large and complex biological systems. Since quantum chemistry at present can only describe the equilibrium state of a limited number of atoms, the introduction of a series of approximations becomes necessary. With large biological systems, the pure *ab initio* approach must be replaced by empirical parameterisation of the model used. These empirical force fields are now routinely used for the simulations of large biomolecular complexes. However, even with such simplified description, it proves challenging to describe large-scale conformational transitions in biomolecules by classical MD simulations. As a consequence, several new methods (long time scale approaches) have been developed. Along with these techniques, some of the main approaches currently used to calculate free energy from MD simulations will be discussed.

Finally, to underpin the concepts of the new computational approach developed in this thesis, a brief outline of the empirical valence bond theory and structure-based potentials is presented.

2.2 Quantum mechanical potentials

The time-independent Schrödinger equation can be used to describe different states of a system where relativistic or time-dependent effects can be neglected [141, 142]

$$\hat{H}\psi = E\psi \tag{2.1}$$

where $\hat{H}$ is the Hamiltonian operator of the system, ψ is the total wavefunction (eigenfunction) and E is the energy (eigenvalue) of the corresponding stationary state of the system. Considering that an analytical solution is only possible for one-electron systems and that most chemical systems include more than one electron, a series of approximations have been introduced. The Born-Oppenheimer approximation assumes that the motion of electrons can be treated separately from that of the nuclei, i.e., the motion of electrons follows the nuclear motion instantaneously. This approximation is justified by the large difference of masses of nuclei and electrons. The total wave function of the system will then be the product of the electronic (ψ_{elec}) and nuclear wavefunctions and the total Hamiltonian operator will be the sum of the electronic

($\hat{H}_{elec}$) and nuclear Hamiltonian operators. If we neglect the nuclear motion in the following, we have a simplified Schrödinger equation for electrons:

$$\hat{H}_{elec}\psi_{elec} = E_{elec}\psi_{elec} \tag{2.2}$$

where E_{elec} is the electronic energy of the system for a given nuclear configuration. It arises from the kinetic energy of the electrons, the electron-electron repulsion potential energy term (V_{ee}) and the nucleus-electron attraction (V_{ne}, the "external" potential acting on the electron). The total energy of the system is the sum of the E_{elec} and the constant nucleus-nucleus repulsion (V_{nn}) terms. The electronic wavefunction of the system (ψ_{elec}) depends on 4n variables (3n space coordinates and n spin coordinates), where n is the number of electrons.

Hartree-Fock (HF) theory approximates the form of the wavefunction ψ_{elec}, by separating its variables, as an antisymmetric product of n individual one-electron spin orbital functions $\rho_i(x)$, each a product of a spatial orbital $\phi_i(r)$, and a spin function. The n one-electron functions can then be used in n eigenvalue equations if we write the electronic Hamiltonian operator as a sum of n one-electron operators. This is possible only in an "independent particle" model where each electron is assumed to move independently in the average potential of all other electrons and nuclei. However, the main deficiency of the HF approximation is the inadequate treatment of the correlation between motions of electrons. Most post Hartree-Fock [143] methods aim to improve on the HF wavefunction and involve electron correlation.

Quantum mechanical calculations can be grouped into *ab initio* and semi-empirical approaches. *Ab initio* calculations do not use any experimental data but only universal physical constants. These methods may provide an accurate description of the electronic structure of a system. However *ab initio* calculations can be extremely expensive in terms of the computer resources required and hence their application to large systems is practically unfeasible. As a consequence, semi-empirical methods were developed to describe large systems based on rigorous quantum mechanical principles, but including several simplifying considerations [144, 145]. These techniques eliminate a vast number of integrals necessary to solve the electronic structure of a system; some are set to constants obtained from experiments or to zero, while others are replaced with simple functions parameterized to reproduce experimental data [146], AM1 [147] PM3 [148] and PM5 [149]. Although the accuracy and performance of semi-empirical methods has

been continuously improved, the application of quantum chemical approaches to simulate the entire biological system of interest is not feasible at present.

2.3 Empirical force fields

Due to the high computational demand required by quantum chemical approaches, empirical potential energy functions (force fields) have been developed to calculate the energy of large molecular systems. The force field of a molecular system is a set of equations with parameters used to approximate certain areas of the *ab initio* potential energy surface (PES) [150, 151]. An optimal electronic structure of the system is assumed in the force field and hence they do not appear in the energy expressions explicitly. Following the Born-Oppenheimer approximation, the potential energy of the system is a function of the atomic positions only. There is a number of force fields that describe the interactions in biological systems surprisingly well, such as the various forms of the AMBER [152, 153], CHARMM [154], and OPLS force fields [155]. Force fields typically describe the molecular system as a collection of beads (atoms) connected by springs (covalent bonds) that obey the laws of classical physics. This represents a fairly good approximation for many biological systems in solution, if electronic changes are negligible. A force field thus contains empirical functions to describe the interaction of bonded and non-bonded atoms. Interactions between bonded atoms are usually described by the bond, angle and dihedral energy terms, while atoms which are three or more atoms apart interact according to van der Waals and electrostatic energy expressions. The total potential energy of the system is then the sum of all the different interactions energy terms:

$$E_{total} = E_{bond} + E_{angle} + E_{dihedral} + E_{vdw} + E_{elec} \tag{2.3}$$

The bond stretching and angle bending terms (1-2 and 1-3 interactions) are usually simple quadratic penalty functions of the deviations of the internal coordinate from a reference value:

$$E_{bond} = \sum k_r (r - r_{eq})^2 \tag{2.4}$$

$$E_{angle} = \sum k_\theta (\theta - \theta_{eq})^2 \tag{2.5}$$

where k_r and k_θ are force constants for bonds and angles with reference values r_{eq} and θ_{eq} and actual values r and θ, respectively. As the harmonic description is valid only for small deformations, some force fields account for the anharmonic effects by adding higher order terms in the potential functions. The torsional potential (1-4 interactions) is a periodic function, which can be described by the leading terms in a Fourier expansion:

$$E_{dihedral} = \sum \frac{1}{2} k_\phi \left[1 + \cos(n\phi - \delta)\right] \tag{2.6}$$

where k_ϕ is a parameter that is proportional to the barrier to rotation, n is the periodicity that indicates the number of minima in the function and δ is a phase angle that determines which torsional angle ϕ correspond to these minima. There is often a need to use special terms for the out-of-plane deformations (improper torsion angles) to ensure that the four atoms in a trigonal plane remain planar.

The van der Waals interactions are approximated by a Lennard-Jones 12-6 potential between pairs of atoms, which account for repulsion at short inter-atomic distances and the long-range attractive dispersion interactions:

$$E_{vdw} = \sum \varepsilon_{ij} \left[\left(\frac{r_{ij}^{eq}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{eq}}{r_{ij}} \right)^{6} \right] \tag{2.7}$$

where r_{ij}^{eq} is the inter-atomic separation for which the energy is at minimum and ε_{ij} is the energy well depth. Electrostatic interactions are typically calculated by Coloumb's law between partial and atomic charges:

$$E_{elec} = \sum \frac{q_i q_j}{\varepsilon r} \tag{2.8}$$

where q_i and q_j are the charges separated by a distance r in a medium with a bulk dielectric constant of ε. Atom-centred charges have clear computational advantages when forces acting on nuclei are calculated. This arrangement however, assumes that the charge density is spherically distributed around atoms in a molecular system and therefore, multipoles are often poorly predicted by these models. Point charges, at least in principle, can reproduce higher electric moments and thus the total electrostatic energy of the system if placed at locations other than the nuclear centres. However, in biological systems there are many charges that depend on the conformation of the molecule and its interactions with other molecules. Consequently, it is not possible to

reproduce the exact electrostatic potential of every configuration using the fixed-charge scheme. Therefore, charges are usually obtained for smaller building blocks of the system and then a "correction" term is applied. This term should account for the changes in charge distribution caused by polarisation of the electrons on each building block in the presence of the remainder of the system. At the present most force fields, such as AMBER, include a polarisable force field [156-158]. As shown in equations (2.7) and (2.8), the interaction energy is inversely proportional to the distance between the interacting particles and asymptotically approaches zero as $r \rightarrow \infty$. On the other hand, the number of these non-bonded terms scales as the square of the number of atoms in the systems. To reduce the computational cost of calculations, a spherical non-bonded cut-off is usually used, beyond which the interaction is set to zero.

Another feature is related to the way boundaries are treated. Due to computational limits, a typical simulated system is generally arranged in a finite-size box. As result a relatively large number of atoms will lie on the surface and will experience different forces from molecules in the bulk. This may seriously affect the results of the simulations and usually, periodic boundary conditions are adopted to reduce the surface effects. In this framework, the system is simulated in a central box surrounded by an infinite number of copies of itself. During the simulation, the molecules in the original box and their periodic images move exactly in the same way. Hence, when a molecule leaves the central box, one of its images will enter the same box through the opposite side. As a result, there are no physical boundaries or surface molecules in the system. It should be noted that the correct treatment of long-range electrostatic interactions is fundamental to preserve the structure, energetics and dynamics of biomolecular systems [159]. As a result, periodic boundary conditions are often used in conjunction with the Particle Mesh Ewald method (PME) introduced by Darden et al. [160]. This approach consists of a fast implementation of the Ewald summation [134], where the total electrostatic energy is divided into two main terms: short-range local interactions and long-range interactions. In the PME framework, short-range interactions are calculated explicitly, while long-range interactions are computed by a summation in the Fourier space providing an efficient method to calculate the total electrostatic energy of the system.

Beyond the functional form of the energy expression, corresponding parameters are crucial to the accuracy of a force field. To limit the number of parameters, atoms with similar chemical environments (same atom-type) share the same parameters. These

parameters can be determined by least-square fitting procedure to minimise the error between calculated and reference data. Often results from *ab initio* quantum mechanical calculations on small molecules are used as the starting point in the parameterisation process. Transferability of the atomic parameters from one system to another and the extrapolation outside regions sampled in the fitting procedure are a basic assumption in the use of a force field. However, the validity of this assumption should be tested by comparing calculated results with experimental data.

Force field development represents a fundamental and ongoing research line. In the present thesis both Amber parm99 [153] and the latest parmbsc0 [152] force fields have been used. In particular the parmbsc0 was introduced to correct the overpopulation of α/γ=(g+/t) backbone angles observed in parm99. Though not used in the present work, it is worth mentioning some recent reparameterisation of the RNA glycosidic torsion as introduced by Banas et al. [161].

2.4 Optimisation of the molecular structure

The potential energy surface (PES) represents the potential energy of the system as a function of the particle coordinates [135]. For a system of N atoms, the PES is a function of 3N Cartesian coordinates. In this context, minimizing the potential energy function means finding the values of the Cartesian coordinates where the energy function is at minimum. A stable molecular structure is usually identified as a minimum on the PES. There may be many minima on the energy surface and the lowest energy minimum is usually referred to as the "*global energy minimum*". Geometry optimisation methods are based on algorithms that aim to locate low-energy points on the energy surface. At a minimum point, the first derivative (gradient) of the energy is zero and the second derivatives (curvature) are all positive. The majority of algorithms are based on the use of derivatives of the potential energy. First order methods use energy gradient (steepest descents or conjugate gradient), while second order methods also use the second derivatives (Newton-Raphson). The minimization algorithm changes the nuclear coordinates in the direction of the forces acting on them toward the energy minimum. The steepest descents method is very efficient far from the minimum and is usually applied to relieve bad contacts or large intra-molecular strains in an initial molecular configuration. The Newton-Raphson method improves the efficiency and convergence to the minimum energy structure. The search for the global energy minimum on the

PES is very difficult and a minimization procedure may never find it. A simple strategy is to monitor the change in energy, coordinates or forces and stop the minimization when the difference between two successive steps is below a pre-determined cut-off. A more efficient way to explore the PES and thus the conformationals space available for a system is to apply molecular dynamics simulation (see Section 2.6).

2.5 Solvent models

A realistic description of the solution environment is critical for the quantitative analysis of biomolecular properties using computer simulations. There are two main approaches to consider the effect of solvents: using explicit solvent models and implicit solvent models. Explicit solvent models include water molecules physically in the simulation. There have been several such water models developed for this purpose: Berendsen's SPC model [162] and Jorgensen's TIP3P [163], TIP4P [163], and TIP5P [164] models are the most generally used. In the present thesis, TIP3P model is used which is based on rigid geometry and static charge approximations. Three atom-centred partial charges are present and are exactly balanced between the positive charges on the hydrogens atoms and the negative charge on the oxygen. Only one van der Waals interactions site is associated with a water molecule and it is localized on the oxygen atom. While this water model is admittedly rather simple, it has been extensively used due to its computational efficiency. However, one limitation of TIP3P is associated to self-diffusion properties, which results in faster dynamics, compared to experimental values for liquid water [165, 166]. There have been several attempts to improve on this simple model by introducing internal flexibility (bond stretching and angle bending) [167], polarization effects [168] or extra interaction sites [163, 164]. Furthermore, explicit inclusion of water molecules in the model increases the number of degrees of freedom in the system. In turn, this leads to slow statistical convergence of molecular properties.

Alternatively, solvent effects can be simulated implicitly as a perturbation to the gas-phase behaviour of the system [169]. These additional equations model the mean-field effect of the solvent and reduce the number of degrees of freedom to be simulated, resulting in faster convergence and lower computational cost. According to implicit models, the solvation free energy can be partitioned in two terms: electrostatic and non-electrostatic term. The non-electrostatic term is given by the contribution of van der

Waals interactions between water and solute molecules and the cavity term derived which is the cost of desolvating the solute. Since both the van der Waals and the cavity terms are mainly related to the first solvation shell, the non-electrostatic part is often approximated with a linear function of the solvent accessible surface area of the solute. There are different ways to calculate the contribution of the electrostatic term. The simplest way is to introduce distance-dependent dielectric function ε(r) in Coloumb's law. However, the most common treatment of electrostatic interactions is based on the use of generalized Born surface area model [170], where the electrostatic term (ΔG_{ele}) is equal to the sum of the Born equation [171] and a term which accounts for the effect of the dielectric medium on the pairwise interactions [170]. In addition, a modified version that incorporates a Debye-Huckel term to account for salt effects at low concentrations has been developed [172]. This approach provides an approximation to the electrostatic term (ΔG_{ele}) that reproduces results of the Poisson Boltzmann (PB) continuum solvent model [173] with increased computational efficiency. Thus, the total solvation free energy (ΔG_{solv}) includes three main terms:

$$\Delta G_{solv} = \Delta G_{ele} + \Delta G_{vdw} + \Delta G_{cav} \tag{2.9}$$

where ΔG_{ele} and ΔG_{vdw} describe the contribution of electrostatic and van der Waals interactions between the solute and the solvent, and ΔG_{cav} is the solute desolvation penalty. One of the limitations of implicit models is their inability to reproduce the microscopic features of the solvent environment.

2.6 Molecular dynamics simulation methods

Computer simulations allow for the exploration and statistical sampling of the potential energy surface and thus the study of both the microscopic and macroscopic behaviour of a molecular system. Macroscopic properties are averages over a representative statistical ensemble of the system. According to the *ergodic hypothesis* the ensemble average of a molecular property can be replaced by its time average [174]. For generating a representative equilibrium ensemble two key methods are available: Monte Carlo [134] and molecular dynamics (MD) simulations [175]. Inherent to the work developed in this thesis, below is briefly described the theory underlying MD simulations.

MD simulations solve Newton's equations of motion for a system of N interacting atoms:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = F_i, \quad i = 1..N^{m_i \partial^2} \tag{2.10}$$

where m_i is the mass of atom i, r indicate the position of atom i, t the time and F the force acting on the atom i. Newton's equations of motion relate the force (F) to the changes in positions as a function of time. The evolution of the atomic coordinates in time is calculated by integrating Newton's equations of motions simultaneously in small time steps. Given initial coordinates and velocities, the forces on the atoms determine the new positions and velocities of all atoms at the subsequent time step. The forces are considered to be constant during the time step and are equal to the negative derivatives of a potential function $V(r_1, r_2, \ldots, r_N)$:

$$F_i = -\frac{\delta V}{\delta r_i} \tag{2.11}$$

Often in order to use longer time steps the highest frequency motions need to be eliminated, for example, by constraining the bond lengths to hydrogens using algorithms such as SHAKE [176], RATTLE [177], or LINCS [178]. As a result, an MD simulation generates a sequence of defined points in the phase space of a system (a trajectory) that are connected in time. In addition MD simulations can be performed to sample from different thermodynamic ensembles. An ensemble is defined as a collection of different microstates, which belong to the same macroscopic or thermodynamic state. Commonly used ensembles are the canonical ensemble (*NVT*), where the number of particles N, the volume V and the temperature T are constant, or the microcanonical ensemble (*NVE*), where N, V and the energy E are constant.

A variant of the classical MD, particularly useful in implicit solvent simulations, is the Langevin dynamics, which is based on the Langevin equations of motion as alternative to Newton's second law. Langevin equation allows mimicking the collision and friction forces, which the solute would experience in the presence of the solvent. For an atom i the Langevin expression can be expressed as follows:

$$m_i \frac{\partial^2 \mathrm{r}_i}{\partial t^2} = \mathrm{F}_i(r) - \zeta \frac{dr}{dt} + R_i(t) \tag{2.12}$$

where ζ_i is the friction coefficient between solute and the fictional solvent, while $R_i(t)$ accounts for the random forces experienced by the atom i. The presence of a friction coefficient ζ_i can improve the conformational sampling in crossing activation barriers with respect to classical MD [175].

In summary, MD simulations can be very effective in exploring the low energy regions of the conformational space. However, one of the main limitations of MD simulations is related to accessible time scales. The affordable simulations times of classical MD range from picoseconds up to a microsecond, although rare cases in the literature have shown millisecond time scale simulations [179]. Biologically relevant processes often happen on the order of milliseconds and longer [12]. In order to study these long time scale events, several so-called enhanced sampling techniques have been introduced in the past years.

2.7 Improved efficiency sampling techniques

To study the biologically relevant time scales by computer simulation techniques is one of the major driving forces behind method development. As mentioned above, the goal of any computer simulation is to sufficiently sample the microstates available to a molecular system. The potential energy landscape represents the conformational space available to a biological macromolecule, discussed extensively in the protein folding community [180-182]. On this landscape, every biomolecular conformation is represented by a specific point. Considering that biomolecules are highly flexible and dynamic in nature, they populate a large number of microstates. At finite temperature in solution, individual molecules constantly hop between different microstates. What prevents a conventional MD simulation to explore the entire energy landscape is the presence of relatively high-energy barriers between the microstates. As a consequence, several approaches have been developed in order to increase the conformational sampling during the simulation. Generally speaking, there are two ways to improve the computational efficiency of standard techniques: through the use of simplified models in terms of their energy function or structural representation or by introducing enhanced sampling techniques to accelerate the conformational sampling. Here I will focus on the latter point and leave the discussion of simplified representations for later.

A possible way to increase the sampling during MD simulations is to raise the simulated temperature above 300 K, thus allowing the system to explore high-energy

regions of the energy landscape. This idea is embodied in simulated annealing [183], where several cycles of increasing and decreasing temperatures are performed successively. This technique is often used in parameter optimisation or structure calculation in NMR spectroscopy [184]. Although accelerated barrier crossing is observed, there is no algorithmic improvement upon classical MD simulations. Another general approach that employs the mean field theory is the so-called Locally Enhanced Sampling (LES) technique, initially introduced by Elber et al. [185]. In this approach, for a given system a region of interest is chosen and multiple copies of that region are generated. While these copies are able to interact with the remaining part of the system, the system will feel only an average force originated from the multiple copies, hence energy barriers are decreased compared to classical molecular dynamics simulations. LES and related mean field methods have been successfully applied to biological problems such as structure prediction [186], free energy calculations [187] or ligand design [188].

Combining the idea of simulating multiple copies of the system, using different temperatures and the Monte Carlo algorithm [134], the replica exchange molecular dynamics technique (REMD) has been developed [189, 190]. In this approach several non-interacting copies (replicas) of the system are simultaneously simulated at different temperatures. Based on a Metropolis-type criterion, the copies can exchange between different temperatures and thus low temperature simulations can escape from local energy minima by jumping to minima sampled by high temperatures simulations. However, for large systems, the required computer time significantly increases, thus limiting the use of REMD for big macromolecular systems. Several successful applications have been reported for the study of peptide and small protein folding [191, 192].

For a system with multiple minima and more than one dominant barrier, the milestone method, developed by Elber and co-workers, could be applied [193]. The main idea is to divide the whole conformational transition path into several smaller steps (milestones) and by combining the transitions between milestones, the entire process can eventually be reconstructed.

Finally, to simulate rare but fast events, which characterize the transition pathway between different conformations, one of the main techniques is the transition path sampling (TPS) [194]. The principal advantage of this approach is that no reaction coordinate has to be defined to describe the transition between the minima. The basic

idea is a generalization of standard Monte Carlo procedure, where conformations are generated randomly but accepted based on a Metropolis criterion. The same type of random walk can be performed in the path space of the transition trajectories and thus generate the transition path ensemble. Given an initial path, algorithms are applied such as the shooting move [195] to generate new trajectories. These trajectories are then accepted or rejected based on an acceptance criterion. Reactive trajectories are defined as those trajectories, which connect the initial and final states.

2.8 Simplified models

Molecular models in which atoms are grouped together and represented by a "pseudo-atom" or "bead" are termed coarse-grained models. These minimal representations coupled with simple energy functions have proved to be very efficient to simulate large-scale conformational changes in biomolecules. These models were pioneered by $G\bar{o}$ [196] and applied to the study of protein folding. Several coarse-grained representations have been proposed in the past for both protein and nucleic acids. Early models were proposed for proteins and consisted of one bead for each amino acid [196-198]. More sophisticated models have also been developed by including two or more beads [199]. Recently, several coarse-grained models have been proposed for both RNA [200, 201] and DNA [202, 203].

These models enormously reduce the degree of complexity of macromolecular systems and allow for simulations of long time scale processes. However, it is important to stress that due to this approximation, exact predictions at the atomic scale are not possible and that these models are best employed to study global motions in large biomolecules. The simplified energy functions associated with the coarse-grained structural representation results in smooth free energy landscapes, where sampling of rare events is significantly accelerated.

One class of simplified energy functions is known as the structure-based or $G\bar{o}$ potential [196]. The basic idea is that the global minimum of this energy function corresponds to a known structure, for example, the native state of a protein. This can be achieved by replacing the non-bonded interactions (van der Waals and electrostatic terms) of classical all-atom force fields with two terms representing the "native" and "non-native" interactions. Usually, native interactions are attractive, while non-native interactions can be considered as repulsive, neutral or less attractive. As a result, the

corresponding potential is a nearly ideal folding “funnel” leading to the native state. Repulsive non-native interactions significantly contribute to smoothing the energy surface with respect to all-atom force fields, by lowering the probability of the system to form metastable states. Structure-based potentials have been shown to be very efficient [201, 204, 205], although they are parameterised for a specific system and thus not transferable.

2.9 Free energy calculations

The main purpose of computer simulations is to relate microscopic states of a system to macroscopic properties such as kinetic and thermodynamic parameters. The most important thermodynamic property of any system is free energy, which is a measure of its stability. Free energy can be calculated as the probability of finding a system in a given microscopic state with respect to the entire phase (or configuration) space accessible to the molecular system. In this case one limitation is the need to sample all the conformational space, which is often not feasible. Several techniques have been proposed to improve the efficiency of sampling the phase space in order to calculate free energy. The approaches described briefly here are free energy perturbation, umbrella sampling and metadynamics. One common feature of these techniques is that they are all dependent on the definition of a specific progress variable or reaction coordinate. The idea of calculating the free energy along a specific coordinate was introduced by Kirkwood in 1935 [206] and is also known as the potential of mean force (PMF).

2.9.1 Free energy perturbation

Free energy is a state function, which means that the free energy difference is only dependent on the initial and final state, no matter what path is taken to go from one to the other. The free energy perturbation technique introduced in 1954 [207] is perfectly suited to study changes in free energy between two separate states. Considering states A and B with corresponding energies E_A and E_B, the free energy difference ΔG can be formulated as follows:

$$\Delta G_{A\to B} = G_B - G_A = -k_B T \ln\left\langle \exp\left(\frac{E_B - E_A}{k_B T}\right)\right\rangle_{\lambda} \tag{2.13}$$

where T is the absolute temperature, k_B is Boltzmann's constant and λ a coupling parameter varying between λ=0 for state A and λ=1 for state B. By simulating the system at state A (λ=0), one can generate an ensemble and calculate the average $\exp\left(-\frac{E_B - E_A}{k_B T}\right)$ from the energy of the initial state A (E_A) and final state B (E_B) for each configuration of the ensemble. One limitation of this approach is that in case whereby the free energy difference between states A and B is larger than $k_B T$, then this estimate may not be accurate enough. In order to check the reliability of the results, the simplest test is to repeat the same procedure by simulating the system at the final state B and verify if there is an overlap in the phase space between A and B. In case there is no adequate overlap between the states, intermediates states may be introduced along λ to improve sampling. In this case the total free energy difference of the process can be calculated as the sum of free energy differences between intermediate states as:

$$\Delta G = \sum_{i=0}^{k=1} \Delta G(\lambda_{i\to i+1}) \tag{2.14}$$

where the interval between λ=0 and λ=1 has been divided into k subintervals.

Common applications of free energy perturbations are the calculation of difference in free energy upon "mutating" a residue [208] or free energy of binding [209].

2.9.2 Umbrella sampling

Umbrella sampling represents one of the most widely used methods to calculate free energy along a predefined reaction coordinate [210]. As discussed above this quantity is also known as the PMF and is derived from the average probability distribution function $\langle p(\xi)\rangle$ of a specific reaction coordinate ξ. However, the direct calculation of $\langle p(\xi)\rangle$ from a single molecular dynamics trajectory is often unfeasible for most biological processes, due to the presence of large energy barriers, which prevent the system from exploring the entire conformational space. The basic idea of the umbrella sampling is to introduce an external biasing potential to adequately sample high-energy regions in the phase space, which is normally only sparsely sampled. By introducing a biasing

potential, the system is forced to sample a small region around a fixed value of ξ (window) and thus generating a more uniform sampling along ξ. The biasing umbrella potential is usually a harmonic function:

$$V_{tot}^{(n)}(q) = V(q) + V_{umb}^{(n)}(q) \tag{2.15}$$

$$V_{umb}^{(n)}(q) = \frac{1}{2} k^{(n)} \left[\xi(q) - \xi_0^{(n)} \right]^2 \tag{2.16}$$

where q is the set of system coordinates, k is the harmonic force constant, and V_{umb} is an umbrella potential that is added to the original system potential V_{tot} to bias the sampling towards a particular value of the reaction coordinate ξ_0. The superscript *(n)* denotes the sampling of a particular value of the reaction coordinate in a series of biased sampling simulations. In order to calculate the free energy, the biased probability distributions need to be combined and unbiased in each window. For this purpose one of the most popular approaches is the weighted histogram analysis method (WHAM) [211]. The WHAM method is a derivation of the multiple histogram method introduced by Ferrenberg and Swendsen [212]. The WHAM algorithm calculates the unbiased probability distribution function as a weighted sum over the data extracted from single distributions, with weights chosen to minimise the variance of the final distribution. Given a number n of biased windows, the biased probability distribution for the window i can be calculated from the umbrella sampling simulation and so the relative biased free energy for that specific value of the reaction coordinate is:

$$G_i^{biased}(\xi) = -k_B T \ln p_i^{biased}(\xi) \tag{2.17}$$

where k_B is the Boltzmann constant. From the biased free energy the unbiased free energy can be calculated:

$$G_i^{unbiased}(\xi) = -k_B T \ln p_i^{biased}(\xi) - V_{umb}(\xi) + F_i \tag{2.18}$$

where V_{umb} is the biasing potential and F_i are the unknown constants calculated through an iterative procedure. Previous studies [213, 214] have shown that the result of WHAM calculations is dependent on the choice of the histogram interval (i.e. bin width). The size of the bin width becomes a trade-off between being small enough so that the probability density does not change significantly within each interval and large enough so to overlap between two successive intervals [215]. Moreover, the optimal bin

width is dependent on the actual values of the harmonic force constant *k* in the biasing potential (Eq. 2.16). As all the approaches herein described, umbrella sampling suffers from the dependence of the chosen reaction coordinate to describe the process. It has been shown that the description of the structural transition along a reaction path, using a given method, is only the first step [216]. In order to understand the mechanism at the molecular level and to identify the transition state ensemble it is essential to define a "*good*" reaction coordinate [216].

An alternative approach to the classic umbrella sampling technique is the adaptive umbrella sampling [217] where instead of using predefined biasing potentials along the reaction coordinate, simulations are carried out iteratively so that the result of one simulation are used to modify the biasing potential for the next simulation.

2.9.3 Metadynamics technique

The Metadynamics technique [218] bring together several features from other techniques and provide a unified method for computing free energy. Although metadynamics, as with many other approaches, rely on the definition of a reaction coordinate, one of its main advantages lies in its ability to treat multiple reactions coordinates simultaneously [218]. The metadynamic algorithm is based on a history dependent random walk, starting from the bottom of the potential well of the energy landscape. In the classical formulation the external biasing potential (V_G) introduced into the system is formed by repulsive Gaussians:

$$V_G(\xi,t) = \omega \sum_{t'=\tau_G,2\tau_G\ldots \atop t'<t} \exp\left(-\sum_{\alpha=1}^{d} \frac{(\xi - \xi_\alpha(t'))^2}{2\delta\xi_\alpha^2}\right) \tag{2.19}$$

where, ω is the height of the Gaussian, $\delta\xi$ is the width of the Gaussian, τ_G the frequency at which the Gaussian functions are added to the potential and α denotes a specific reaction coordinate out of the total number of reaction coordinates *d*. These parameters are very important to adjust for the speed and accuracy of the simulation. In particular, wide Gaussians will allow a fast exploration of the phase space, but with large errors in the resulting free energy. On the other hand, narrow Gaussians deposited infrequently will increase the accuracy of the calculation, at the expense of increased simulation time. The basic assumption in metadynamics simulation is that given sufficient time, the sum of the deposited Gaussians provides an estimate of the underlying free energy:

$$\lim_{t\to\infty} V_G(\xi,t) \approx -G(\xi) \tag{2.20}$$

where V_G is the resulting potential recalculated from the sum of all the deposited Gaussians and G the final free energy. When multiple reaction coordinates are used, the number of Gaussian functions necessary to escape from the minimum will be proportional to $(1/\delta\xi)^d$ [218]. Furthermore, often the use of many reaction coordinates would be computationally prohibitive. As a result, some improvements have been proposed recently toward a better sampling of the energy landscape. Some variants of the standard metadynamics include the "multiple walkers" [219], where multiple simulations are running simultaneously and all contributing to the same metadynamics bias. Parallel tempering metadynamics [220] is based on running different copies of the system at different temperatures.

2.10 Empirical valence bond theory

The empirical valence bond (EVB) theory was introduced by Warshel in 1980 [221]. This technique was designed as a consistent way of transferring gas phase charges and potential energy surfaces to solutions and protein interiors. As in conventional QM/MM calculations [222], the system of interest is formed of a central reactive region where the electronic changes take place and the surrounding region, which interact with the central region through van der Waals, electrostatic and bonding interactions.

The EVB method is based on the valence bond (VB) theory which describes a chemical reaction path in terms of resonance structures or more precisely diabatic states corresponding to valence bond structures. The most important valence bond configurations for a given reaction have a clear physical meaning and hence, it is possible to interpret different structures (e.g. reactant and products) in terms of these valence bond states. As an example, the proton-transfer reaction can be considered: AH + B → A^- + B^+H. According to the VB theory, three resonance structures can be defined: 1) A—H B; 2) A^- H^+ B; 3) A^- B^+—H. However, if the highest energy structure (A^- H^+ B) is neglected, such a reaction can be reduced to an effective two state problem, where in the initial state the proton is bound to atom A and in the final state is bound to atom B. The matrix elements, that define the potential energy of each state, can be evaluated by using classical Quantum mechanical approaches. The novelty introduced in the EVB model, consists of simplifying the analytical form of these states by using

empirical force field energy functions. For a system with two VB states V_{11} and V_{22}, the total Hamiltonian is given by:

$$\begin{aligned} V_{11} &= V_1 \\ V_{22} &= V_2 + \Delta\alpha_{12} \\ H &= \begin{vmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{vmatrix} \end{aligned} \tag{2.22}$$

where the matrix elements V_{11} and V_{22} are simply the potential energies of the reactant and product states. The coupling elements *(V_{12}=V_{21})* describe the physics needed for transitions between the diabatic states. The off-diagonal elements (V_{12}) of the EVB Hamiltonian are represented by simple functions (usually exponential or Gaussian) of the solute coordinates, or in some cases constant values might be used [223]. In addition the diabatic state shift ($\Delta\alpha_{12}$) is introduced to adjust the relative energy of each basin. The total potential energy (E_0) of the system can be obtained by solving the characteristic equation:

$$H_0 C_0 = E_0 C_0 \tag{2.21}$$

where C_0 is the ground state coefficient vector, E_0 is the corresponding eigenvalue and H_0 the hamiltonian. The reactive energy surface is obtained as the lowest adiabatic surface by diagonalizing the Hamiltonian matrix, which is formed of non-reactive diabatic states. The solution to this two-state case has the following analytical form:

$$E_0 = 0.5\left(V_{11} + V_{22}\right) - \sqrt{\left[0.5\left(V_{11} - V_{22}\right)\right]^2 + V_{12}^2} \tag{2.23}$$

A key feature of the EVB method is the possibility to calibrate the energy surface to a high degree of accuracy against available experimental or quantum mechanical calculations, by tuning the off-diagonal element (V_{12}) and the diabatic state shift ($\Delta\alpha_{12}$). The chemical process can also be represented by more than two (*VB*) states, including possible intermediate states. EVB is mostly used to describe chemical reactions rather than conformational changes. The major advantage of the EVB method lies in its computational efficiency since diabatic energy surfaces are calculated by simple analytical functions.

2.11 Summary

In the last decades, the application of computational approaches to study biological problems has grown exponentially. Theoretical methods became a powerful tool that can be applied to study mechanistic aspects of biology at the atomic level. In particular, these methodologies have been proven very useful in complementing experimental results. However, direct molecular dynamics simulations are often limited in describing long time-scale processes. Enhanced sampling techniques have thus been developed to speed up calculation and to simulate large-scale conformational changes in biomolecules. In addition, one of the main goals of computer simulations is to relate microscopic properties of a system to macroscopic observables: kinetic and thermodynamic data. Also, they can provide a theoretical interpretation to single molecule experiments. However, no universal technique exists at present and different approaches can be applied for different purposes.

In this thesis work a new computational tool, for the study of conformational changes in biomolecules, has been developed, by combining different features of theoretical models presented here.

Chapter 3

Development of a novel computational technique

3.1 Introduction

Function of biological macromolecules is inherently linked to their complex conformational behaviour. Many biological processes involve large-scale structural changes, often characterized by several intermediate conformations between the initial and final states. These structural changes are usually the key to understanding the mechanism underlying the function of many essential biomolecular machineries. Structural data on the initial, final and intermediate conformations can generally be obtained by experimental techniques, such as X-ray crystallography or NMR spectroscopy (discussed in Section 1.4). It remains, however, challenging to describe the mechanism of conformational transition between the various stable and metastable states and their associated dynamic behaviour. Molecular simulation can, in principle, account for the structural flexibility of individual states and bring these together to rationalize the overall process. Nevertheless, conventional simulation techniques are currently not capable of describing complex conformational transitions that typically occur on timescales ranging from milliseconds to seconds.

Several approaches have been proposed to overcome the time-scale limit by coarse-graining the molecular structure for proteins [224] and more recently for nucleic acids [200, 225]. Alternatively, simplified energy functions may be used in conjunction with atomistic models [201, 226, 227]. The main purpose of using minimalist approaches is to reduce the complexity of the system, which, in turn, allows simulations or much longer timescales to be carried out. It has been shown that the native topology plays a key role in the folding process of small proteins [228] and that the folding rate is proportional to the contact order [229]. Moreover, it is now accepted that these smaller

systems evolved to fold without significant frustration to their respective native structures [230, 231]. Simple and tuneable potentials can be devised based on the contacts present in the native structure, which result in funnel-like energy surfaces around a stable structure. This approach has been successfully applied to study protein folding [232, 233]. More recently, similar ideas have also been explored for nucleic acids [201, 234].

Structure-based potentials are typically used to describe motions around a single dominant minimum. Recently however, several attempts have been made to use such simple potentials to construct energy surfaces with multiple minima for studying conformational transitions in biomolecules. Zuckerman introduced a "doubly-native" generalisation of a coarse-grained structure-based potential [216, 235]. Best et al. developed a combined potential from two separate structure-based potentials using an exponential Boltzmann weighting scheme [204]. Elastic networks have been merged using the empirical valence bond (EVB) approach [221] to study conformational changes [236, 237]. A similar approach was used to combine coarse-grained structure-based potentials to describe conformational transition in proteins [205].

In this chapter a novel method will be presented [238] that extends earlier methods to simulate conformational transitions efficiently in large biomolecular systems. This approach uses atomistic structure-based potentials (SBP) to describe individual conformational states, which are then coupled by the EVB theory to define a unified multiple-basin energy landscape. This method, termed EVB-SBP, exploits structural and energetic information available from experimental data and seeks to describe biological processes at the atomic level. In the following sections, the theory, implementation and testing of this novel approach are described.

3.2 Potential energy function

Structure-based potentials (SBP) are employed to describe the potential energy landscape that corresponds to individual conformations. The energy function of a classical all-atom force field can be divided into two main components: the 'bonded' part that includes covalent bonds, angles and dihedrals, and the 'non-bonded' part that comes from the combined effect of van der Waals and electrostatic interactions (see Section 2.3). It is well known that the potential energy landscape defined by all-atom force fields is very "rough", with several local minima corresponding to meta-stable

states and an equal number of energy barriers between those [239]. Furthermore, the evaluation of the non-bonded term in the potential energy function is computationally the most expensive part of molecular simulations. SBP may provide an optimal solution to this problem by greatly simplifying the non-bonded energy term, which results in an improved computational efficiency as well as a smoother energy landscape of the system.

In the following, the methodology used to generate SBP from stable structures is described. Experimental structures (X-ray or NMR) are usually employed as a starting point, although three-dimensional models generated by homology modelling or other specialised methods can also be utilized. The basic idea is to exploit a 'reference' molecular structure to generate a matrix of inter-atomic distances encoding the three-dimensional information. No modification is applied to the interaction potentials for covalent bonds, angles and dihedrals, which are taken from the Amber force field for nucleic acids (parm99 + parmbsc0) [152, 153] (see Chapter 2 Eq. 2.4-2.6). The non-bonded term instead is modified substituting van der Waals and Coulomb interactions with two terms named *native* and *non-native* interactions. Native interactions are responsible to reproduce local inter-atomic contacts in the reference structure, while non-native interactions ensure that atoms distant in the reference structure will keep away from one another. The total energy function, for a Lennard-Jones 12-6 potential representing native interactions, has the following form:

$$V = \sum_{bond} K_r (r - r_{eq})^2 + \sum_{angle} K_\theta (\theta - \theta_{eq})^2 + \frac{1}{2} \sum_{dihedral} K_\varphi \left[1 + \cos\left(n\varphi - \gamma\right)\right] + \sum_{native} S\varepsilon \left[\left(\frac{r_{ij}^{eq}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{eq}}{r_{ij}}\right)^{6}\right] + \sum_{non-native} \varepsilon \left(\frac{r_{ij}^{eq}}{r_{ij}}\right)^{12} \quad (3.1)$$

where r is bond length, θ is bond angle and φ is dihedral angle between covalently bonded atoms with *eq* indicating the corresponding equilibrium values and K the force constants. In the dihedral term n and γ represent the usual multiplicity and phase, respectively. It has to be noted that in the parm99 force field for each dihedral interaction 1-4 a van der Waals and an electrostatic component are calculated separately from non-bonded energy terms and are scaled by a factor of 2 and 1.2 respectively. In this model, since no electrostatic term is included, only repulsive 1-4 van der Waals interactions are calculated. In the remaining non-bonded interactions a Lennard-Jones

(LJ) 12-6 potential is calculated where ε is the well depth and S is a scaling factor that was introduced to calibrate the strength of native interactions. Native and non-native contacts are defined on the basis of atomic distances in the reference structure. Interactions between non-hydrogen atom pairs whose distances are within a cut-off value (Fig. 3.1) and belong to different residues are considered to be in native contact, while all the others are considered non-native.

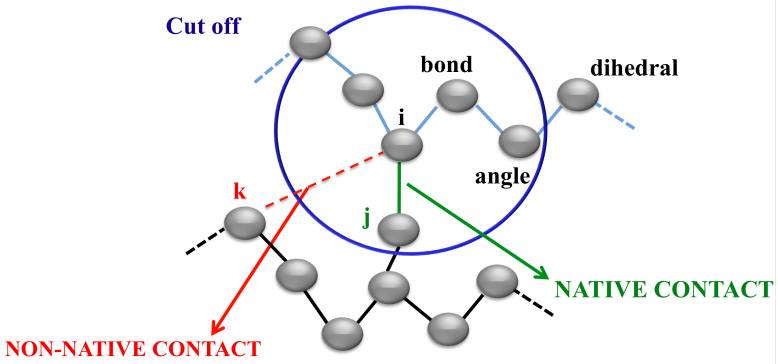

Figure 3.1 Definition of native and non-native contacts. The blue circle represents the cut-off distance around atom *i*. The contact between atom *i* and non-bonded atom *j* within the cut-off is considered native, while the contact with non-bonded atom *k* that lies outside the cut-off is considered non-native.

The choice of the cut-off is a trade-off between an accurate description of the native structure and an increased computational efficiency. A high cut-off value increases the total number of interactions which define the native state, and hence the accuracy, but also decrease the computational efficiency due to a larger number of native contacts that has to be evaluated in the total energy function. In addition, an optimal cut-off should not be too long, as that would effectively eliminate the difference between the native contact lists of alternative conformations. In our model, each native contact is described by the classical Lennard-Jones potential (Eq 3.1), depicted in Figure 3.2. The Lennard-Jones potential is simply the sum of the attractive and repulsive components (Fig. 3.2a), where the latter component also serves to describe repulsive non-native interactions. In addition, using the rescaling factor S one can modulate the strength of the native interactions (Fig. 3.2b).

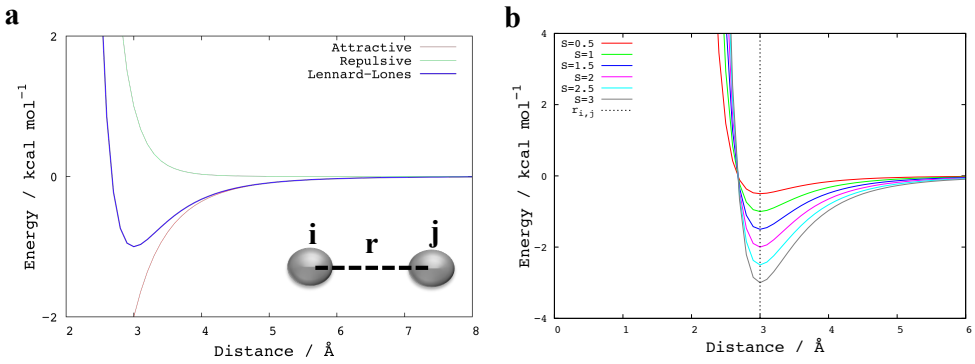

Figure 3.2 Lennard-Jones potential between two arbitrary atoms in the system, *i* and *j*. (a) Repulsive (green) and attractive (red) components of the LJ potential (blue). (b) Effect of changing a rescaling factor (S) on the LJ potential with r_{ij}=3Å and ε=1 kcal mol^{-1}.

Alternatively, different potential functions, such as harmonic or Morse, are equally applicable to describe native interactions. The harmonic potential in the classical form is:

$$V = k(r_{ij} - r_{ij}^{eq})^2 \tag{3.2}$$

where k is the harmonic force constant, r_{ij} and r_{ij}^{eq} respectively the observed distance and the equilibrium distance between atoms i and j. Using an harmonic potential, the atoms i and j will fluctuate around the minimum and will experience an energy penalty, which is proportional to the displacement from the equilibrium distance (r_{ij}^{eq}). In addition, due to the symmetry of the potential, the penalty for a given displacement will be the same either increasing or decreasing the *i-j* distance. As a result, this potential is suitable to reproduce conformations near the equilibrium geometries, but fails in describing systems far from the minimum. On the other hand, the Morse potential has the following form:

$$V = D_e\left[1 - e^{-\alpha(r_{ij} - r_{ij}^{eq})}\right]^2 \tag{3.3}$$

where D_e is the well depth, α controls the curvature of the potential around the minimum and r_{ij} and r_{ij}^{eq} are the observed and equilibrium distances between atoms i and j, respectively. As with the Lennard Jones potential, the Morse is an anharmonic function, including a short-range repulsive part and a long-range attractive term. Both these potentials are usually parameterised to reproduce experimental or ab initio calculations. However, the Lennard Jones function is mostly used to describe non-

bonded interactions. In this work, harmonic, Morse and Lennard Jones potentials have all been implemented and tested on an elementary system (see below), while the successive applications (Chapters 4 and 5) of this method will employ the Lennard Jones potential.

Given a Lennard Jones potential to describe a native interaction between atoms *i* and *j*, the minimum of the interaction potential can be considered as corresponding to the exact distance between those atoms in the reference structure. In addition, an alternative approach has also been used in this work for one application (see Chapter 4): r_{ij}^{eq} is taken as the combined van der Waals radii from the Amber parm99 force field. In this case, atom pairs of a given "atom-type" combination, as described in the Amber force field, will exhibit an energy minimum at the same position of r_{ij}, which should result in an even smoother energy surface. Such an approach is expected to reproduce reference structures relatively well if we consider that a cut-off used for native interactions of about 4 Å is close to the combined van der Waals radii of most atom pairs in the Amber force field. It is worth noting that the choice of a cut-off value and a rescaling factor (*S*) is crucial to determine both structural stability and dynamical fluctuations of the molecular system. While the choice of the cut-off can be made *a priori*, the rescaling factor requires parameterisation against data from all-atom force field simulations or experimental data for atomic fluctuations, such as the crystallographic B-factors. The optimal rescaling factor can be obtained by performing simulation with different *S* values and evaluate the correspondence to reference structural or energetic data (see Chapter 4.). An alternative procedure to introduce atom pair specific well depth has also been developed (see Section 3.6).

In summary, the structure-based potential described above has the capacity to provide a funnel-shaped potential energy surface where random configurations of the system will minimise their energy toward the global minimum, the native state. Repulsive non-native interactions play a key role in smoothing the energy surface because these prevent the formation of metastable states, which are kinetic traps on the folding pathway. Repulsive non-native interactions can however preclude the formation of stabilising non-native interactions in, for example, the transition states. Such interactions have however been shown to play only limited roles in proteins [240, 241]. The use of SBP in a simulation is thus in agreement with the folding theory of minimal frustration [230, 231, 242], in which biomolecules fold through sequential steps of partially folded states, stabilised by native interactions, toward the final folded state.

3.3 Coupled structure-based potentials

The atomistic structure-based potentials described above can be used to define a smooth energy surface without losing the chemically relevant details, but it would encode only for one dominant minimum in the potential energy landscape. Since many biological processes involve several stable intermediate conformations between the initial and the final states, this energy function is unable to describe such processes itself. To simulate structural transitions between separate energy basins, a combined potential energy surface must first be defined. For this purpose, the EVB theory of Warshel [221] was used with a key difference from the original model: each diabatic state in the Hamiltonian is represented by a unique structure-based potential related to a single conformation (V_i).

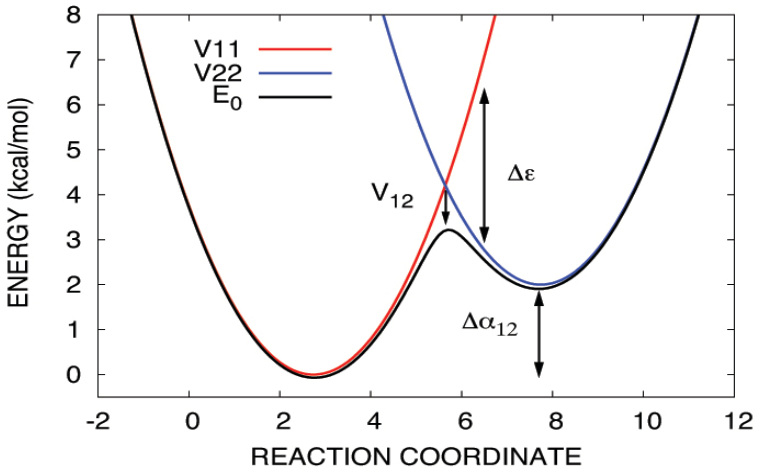

Figure 3.3 Coupling of diabatic states (V_{11} and V_{22}) representing individual conformations by using the EVB theory to give a unified potential energy surface (E_0). Arrows indicate the effect of coupling element (V_{12}), diabatic state shift ($\Delta\alpha_{12}$) and the energy gap reaction coordinate ($\Delta\varepsilon$).

Two parameters were employed to construct the combined potential energy surface: (i) diabatic shift ($\Delta\alpha_{ij}$) that determines the relative stabilities of basins *i* and *j*, and (ii) coupling element (the off-diagonal elements of the symmetric Hamiltonian matrix) that determines the shape of the surface at the crossing of the diabatic potentials (Fig 3.3). The latter is also important to provide the correct energy barrier for the conformational transition and to obtain a smooth transition state region that is mathematically differentiable. The combined multiple-basin potential energy surface is obtained as the lowest adiabatic surface by diagonalizing the Hamiltonian matrix. Thus the EVB

method can be used to construct an energy surface that reproduces the required energetics of conformational transitions using simple structure-based potentials.

Obviously, these potentials describe structures best in the proximity of the reference state and the transition region is obtained through extrapolation. Nevertheless, the major advantage of this scheme lies in its computational efficiency, which allows for long time-scale simulation of biomolecules to be carried out with moderate computational resources.

When a reactive event is described by a high free energy barrier, standard molecular dynamics on the EVB ground-state surface will not adequately sample the important transition-state region. Under these conditions, transitions are rare events and sampling on the EVB surface effectively reduces to sampling on a diabatic surface. Biased simulations were employed to enhance sampling along a given reaction coordinate, using the umbrella sampling technique (see Section 2.9.2). The reaction coordinate used here, by analogy with electron transfer reactions [201], is defined as the energy gap ($\Delta\varepsilon$) between the potential energies of the diabatic states at a given structure: $\Delta\varepsilon=V_{11}-V_{22}$. This general reaction coordinate describes changes in the entire system without resorting to simple geometric progress variables.

3.4 Implementation of the EVB-SBP method

The basis for the EVB-SBP code development is the `sander` module in the Amber molecular dynamics simulation package version 10 written in Fortran 77 language [243]. This software includes the parallel "multisander" architecture introduced via the message passing interface (MPI) implementation. The classical empirical valence bond approach has been implemented on the top of the multisander infrastructure. Thus, information for each EVB diabatic state is obtained from separate (simultaneous) instances of `sander`. In the following, a brief description of the main EVB subroutines and their functions is provided. These can be divided into three main groups according to the EVB procedure: 1) initialisation of the EVB method (*evb_vars; evb_input*); 2) calculation of the diabatic state energies (*morsify; mod_vdw; evb_ntfrc*); 3) calculation of the EVB ground state energy and corresponding forces (*evb_force*). All the variables, constants and arrays, related to the EVB method are declared in the subroutine *evb_vars* (`sander` code file: evb_vars.f). The subroutine *evb_input* (`sander` code file: evb_input.f) is involved in processing the EVB input files. The subroutine *morsify*

(`sander` code file: morsify.f) calculates the native energy contribution for each diabatic state as a sum over the whole atom pair list, defined in the input file. The subroutine *mod_vdw* (`sander` code file: mod_vdw.f) can be used to exclude van der Waals interactions between specified atom pairs. The main interface of the EVB is represented by the subroutine *evb_ntfrc* (`sander` code file: evb_ntfrc.f), where both the bonded and non-bonded energy terms are combined to calculate the energy of the diabatic state. The energies and forces of the diabatic states are communicated via MPI to the master node, which is responsible for computing the EVB ground-state energy and corresponding forces (subroutine *evb_force*), and subsequently broadcasting these back to the slave nodes for the next molecular dynamics step.

Furthermore, the umbrella sampling implementation along the energy gap reaction coordinate is herein described. The main difference with respect to unbiased simulations is related to the calculation of the total energy of the system and the corresponding forces (subroutine *evb_force*). The total energy of the system (V_{tot}) is calculated as the sum of the EVB ground state energy (V) and the umbrella potential (V_{umb}) introduced at fixed values (umbrella windows) of the energy gap reaction coordinate (Eq. 2.15-2.16). For a given structure, the energy penalty (V_{umb}) is proportional to the displacement of the calculated energy gap ($\Delta\varepsilon$), with respect to the reference value ($\Delta\varepsilon^{(n)}$) in the corresponding umbrella window. Resulting forces are directly proportional to the umbrella potential, and act on the system by favouring the sampling of the conformational space around the reference values ($\Delta\varepsilon_0^{(n)}$). The following figure shows the `sander` code (subroutine *evb_force*) where the umbrella sampling along the energy gap reaction coordinate is actually implemented (Fig. 3.4).

```
!  Umbrella Sampling along an energy gap reaction coordinate

   case( "egap_umb" )

    egapRC = .true.

    do n = 1, nbias

!   ni and nf are the initial and final states.
    ni = bias_ndx(n,1)
    nf = bias_ndx(n,2)

!      Calculate energy gap reaction coordinate
       RC = evb_Hmat%evb_mat(ni,ni) - evb_Hmat%evb_mat(nf,nf)
       evb_bias%RC(n) = RC

!      Calculate Umbrella potential
!      k_umb(n) is the umbrella force constant
!      r0_umb(n) is the energy gap reference value
       evb_bias%nrg_bias(n) = 0.50d0 * k_umb(n) * ( RC - r0_umb(n) )**2

!      Calculate Umbrella forces
!      xf(:,ni) is equal to dV11/dq
!      xf(:,nf) is equal to dV22/dq
       evb_bias%evb_fbias(:,n) = k_umb(n) * ( RC - r0_umb(n) ) &
                                    * ( xf(:,ni) - xf(:,nf) )
    enddo

    do n = 1, nbias

!      Update total forces
       evb_frc%evb_f(:) = evb_frc%evb_f(:) + evb_bias%evb_fbias(:,n)

!      Update total energy
       evb_frc%evb_nrg  = evb_frc%evb_nrg + evb_bias%nrg_bias(n)

    enddo
```

Figure 3.4 File *evb_force*.f of the sander code, where umbrella sampling along an energy gap reaction coordinate is implemented.

For a given structure, the energy gap is calculated from the diabatic state energies and the umbrella potential is computed. The resulting forces are calculated as the negative derivative of the umbrella potential with respect to the atomic coordinates:

$$F = -\frac{dV_{umb}^{(n)}}{dq} = -k^{(n)} * \left(\Delta\varepsilon - \Delta\varepsilon_0^{(n)}\right) * \frac{d\Delta\varepsilon}{dq} = -k^{(n)} * \left(\Delta\varepsilon - \Delta\varepsilon_0^{(n)}\right) * \frac{dV_{11}}{dq} - \frac{dV_{22}}{dq} \qquad (3.4)$$

where q is the set of atomic coordinates, $k^{(n)}$ the umbrella force constant, n denotes a particular value of the reaction coordinate, $\Delta\varepsilon$ is the energy gap for a given set of atomic coordinates and $\Delta\varepsilon_0^{(n)}$ is the energy gap reference value. The derivative $d\Delta\varepsilon/dq$ is calculated as the derivative of the difference between the diabatic state potential V_{11} and V_{22} with respect to the atomic coordinates. It is important to note that the derivative dV_{11}/dq is equal to the sum of the forces calculated for each component of the diabatic

state potential (Eq. 3.1). The same argument applies to dV_{22}/dq. Finally, both umbrella potential and umbrella forces are added respectively to the EVB ground state energy and to the EVB forces in order to obtain the total potential energy of the system and the total forces.

The source code was herein modified to combine the EVB method with customised structure-based potentials. A flow diagram showing the connection between the files which have been modified is depicted in Figure 3.5.

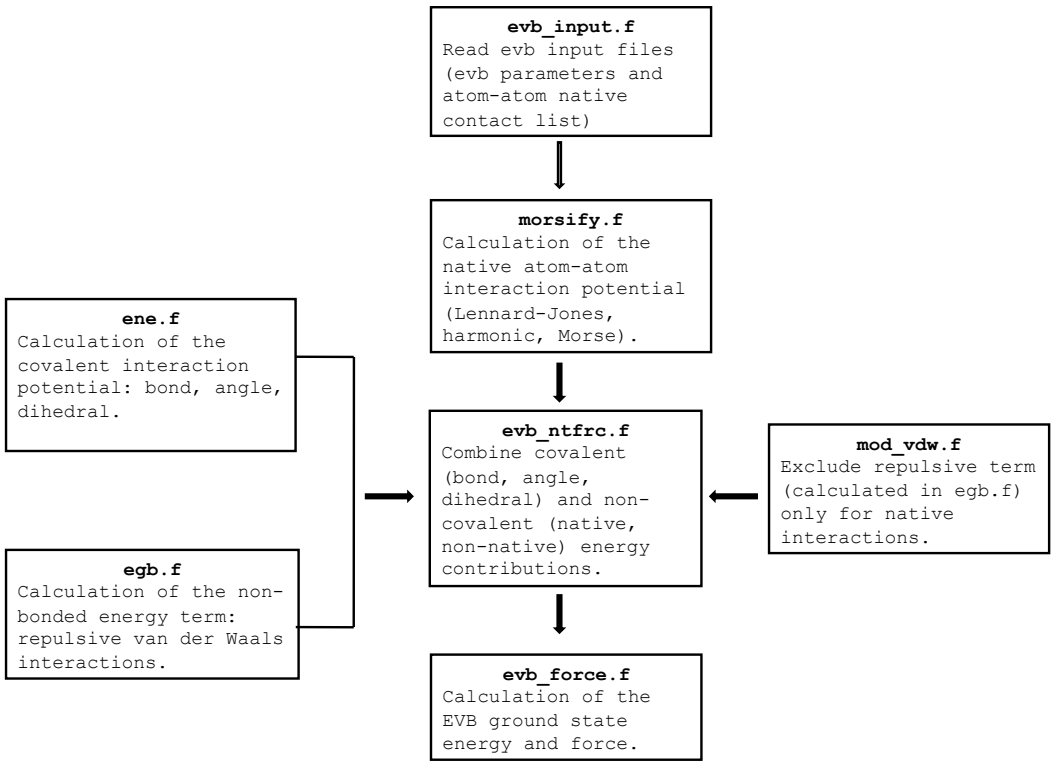

Figure 3.5 Data flow in the modified `sander` code [243].

The steps to be followed in order to implement the EVB-SBP method can be summarized as follows. The first step is to introduce the desired potential energy function for native interactions. In the original implementation of the EVB method, the Morse potential [244] is used. As described above, for the present calculations the classical 12-6 LJ potential was employed, although Morse and harmonic potentials have also been used for testing purposes. Consequently, the subroutine *morsify* was modified to introduce the new energy function and the corresponding derivatives and to update the total force array. The following illustrates how the form of the original Morse potential (Fig 3.6a) can be changed with any other desired classical potential (e.g.

Lennard-Jones or harmonic). The subroutine *morsify* will operate on a list of native atom-atom interactions (defined in the EVB input file) specific to a given conformation or diabatic state in the EVB theory. The calculated native energy will contribute to the potential energy of the EVB state.

A

```
! Variables
  D  = morse(n)%D   ! Well Depth
  a  = morse(n)%a   ! Parameter to control potential curvature
  r0 = morse(n)%r0  ! Equilibrium distance

! Calculate Morse Potential
  exp_part = exp( - a * ( rij - r0 ))
  vmorse(bead_dx) = vmorse(bead_dx) + D*(1.0d0 - exp_part)**2

! Calculate derivatives
  ff = 2.0 * D * (1.0d0 - exp_part) * a * exp_part / rij
```

B

```
! Variables
  a  = morse(n)%a  ! Harmonic force constant
  r0 = morse(n)%r0 ! Equilibrium distance

! Calculate harmonic potential
  delr2inv = 1 / rij
  rdst = rij - r0
  vmorse(bead_dx) = vmorse(bead_dx) + a * rdst**2

! Calculate derivatives
  ff = 2 * a * rdst * delr2inv
```

C

```
! Variables
  D  = morse(n)%D   ! Well depth
  r0 = morse(n)%r0  ! Equilibrium distance

! Calculate Lennard Jones potential
  delr2inv = 1 / (rij**2)
  r6  = delr2inv * delr2inv * delr2inv
  r6_2 = (r0**2) * (r0**2) * (r0**2)
  f6  = 2.0 * r6_2 * r6
  f12 =  (r6_2 * r6_2) * (r6 * r6)
  vmorse(bead_dx) = vmorse(bead_dx) + D*(f12 - f6)

! Calculate derivatives
  ff = -D * (12.d0*f12 - 6.d0*f6) * delr2inv
```

Figure 3.6 Modification to the file morsify.f of the `sander` code are shown: a) original Morse potential; b) harmonic potential; c) Lennard-Jones potential.

The other fundamental contribution is coming from non-native interactions, here considered as repulsive. For this purpose the attractive part of the van der Waals potential, has been removed by modifying the subroutine *egb* (`sander` code file: egb.f) (Fig. 3.5), in order to calculate only the repulsive term for non-bonded interactions. However, the repulsive term is also calculated for the native interactions. As a consequence, the subroutine *mod_vdw* was modified to subtract the van der Waals repulsive term calculated for the native atom pair list. In addition, no electrostatic contribution to the final potential energy function is considered. For this purpose the subroutine *egb* was also modified to avoid electrostatic energy calculation.

Finally, the adiabatic ground state energy of the system is calculated from the energies of individual diabatic states.

One last clarification is necessary and it is related to the definition of the minimum (r_{ij}^{eq}) of the Lennard-Jones potential. As mentioned above, there are two ways of defining these parameters: consider r_{ij}^{eq} as the interatomic distance between atom pair *i-j* in the reference structure or, alternatively, as equal to the combined van der Waals radii of atoms *i* and *j* defined in the Amber parm99 force field [153]. In this latter case, a convenient way of implementation is to read directly these parameters from the topology file where this information is readily available (Fig 3.8). In the topology file, two parameters named as A and B coefficients can be found for each non-bonded interaction pair. A is equal to $\varepsilon(r_{ij}^{eq})^{12}$, while B is $2\varepsilon(r_{ij}^{eq})^{6}$ and the resulting interatomic potential is equal to:

$$V_{ij} = S\left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}\right) \tag{3.5}$$

where *S* is the rescaling factor. In the case that the positions of the minima are defined as the exact inter-atomic distances in the reference structure, native contacts should be divided into two groups [205]: contacts that are present in only one of the reference structures and contacts that are present in both reference structures. A particular treatment has to be applied when a native contact is present in different reference structures with different energy minima. Suppose that the native distance r_{ij} is present in two reference structures A and B, where $r_{ij\text{-}A} < r_{ij\text{-}B}$ and that the system is in state A where the inter-atomic distance *i-j* is approximately $\sim r_{ij\text{-}A}$. In this case, while the native energy contribution for the *i-j* contact is favourable in state A, it is highly repulsive in state B at the same geometry (Fig. 3.7). Since the expression for the EVB ground state energy involves the energy difference between the diabatic states at a given geometry (see Eq. 2.23), this will lead to unreasonably large energy differences for the distance change of that *i-j* atom pair. This energy will contribute significantly to the sum of all the atom-atom pair interactions defined in the native contact list corresponding to that state. Furthermore, the energy gap between the diabatic states is used as the reaction coordinate to drive the conformational change, and large fluctuation in the reaction coordinates would also lead to numerical instabilities in the calculations.

One possible solution to avoid this problem is to introduce a flat bottom potential between the two minima $r_{ij\text{-}A}$ and $r_{ij\text{-}B}$ (Fig 3.7) [203].

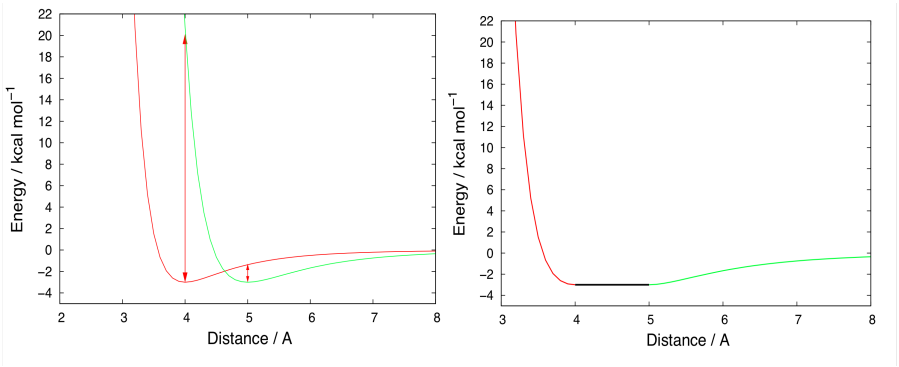

Figure 3.7 Example of a native contact present in both reference structures with different equilibrium values. (left) Two separate 12-6 LJ potentials results in high energy gap between EVB diabatic states (right); introduction of a flat bottom potential between the minima of the original LJ potential alleviates the problem.

In the following, the modifications to the subroutine *morsify* are shown for different definitions of the native Lennard-Jones potential minimum.

A

```
! Calculate index (ic) for  A (cn1) and B (cn2)
! coefficients arrays, as described in the Amber
! topology format.

    iaci = ntypes * (iac(i) - 1)
    ic = ico( iaci + iac(j) )

! Rescale A and B coefficient by a factor s

    rcn1 = cn1(ic) * s
    rcn2 = cn2(ic) * s

! Calculate Lennard-Jones potential

    delr2inv = 1 / (rij**2)
    r6  = delr2inv * delr2inv * delr2inv
    f6 =  rcn2 * r6
    f12 = rcn1 * r6 * r6
    vmorse(bead_dx) = vmorse(bead_dx) + (f12 - f6)

! Calculate derivative
    ff = (-12.d0 * f12 + 6.d0 * f6) * delr2inv
```

B

```
! Variables
  D  = morse(n)%D   ! Well depth
  a  = morse(n)%a   ! Equilibrium distance (r_ij^eqA)
  r0 = morse(n)%r0  ! Equilibrium distance (r_ij^eqB)

!  Calculate potential
  f0 = 1.0d0

  if (rij <= a) then
  r6_2 = (a**2) * (a**2) * (a**2)
  f6  = 2.0 * r6_2 * r6
  f12 = (r6_2 * r6_2) * (r6 * r6)
  else if (rij < r0) then
  f6  = 2.0d0
  f12 = 1.0d0
  f0 = 0.0d0
  else if (rij >= r0) then
  r6_2 = (r0**2) * (r0**2) * (r0**2)
  f6  = 2.0 * r6_2 * r6
  f12 = (r6_2 * r6_2) * (r6 * r6)
  end if

  vmorse(bead_dx) = vmorse(bead_dx) + D * (f12 - f6)

! Calculate derivatives
  ff = -D * (12.d0 * f12 - 6.d0 * f6) * delr2inv * f0
```

Figure 3.8 The subroutine morsify.f has been modified to obtain r_{ij}^{eq} as the sum of van der Waals radii of atoms *i* and *j* defined in the amber parm99 force field [153](A), or alternatively as the interatomic distance between atom pair *ij* in the reference structure, with the introduction of a flat bottom potential (B).

3.5 A simple test model

In order to test the new numerical method, a simple model system was established with two interacting point-mass particles. In the initial state, the inter-particle distance is ~2.74 Å and in the final state the distance is increased to ~7.74 Å (Fig. 3.9). As a first step, an energy surface with two minima was defined using the EVB coupled structure-based potentials. This simple model was used to gain a good understanding of the new method. There is only one native contact to be introduced between the two particles. The low-dimensional potential energy surface, thus defined, ensures that forward and backward trajectories are fully converged in order to calculate the underlying free energy. The results of the numerical simulation can also be compared with those calculated from analytical expressions. Further, using this test system, it is easy to follow the relationship between the non-geometric reaction coordinate (energy gap) and the single distance between the two particles. The effect of varying the EVB parameters, the diabatic shift and the coupling element, on the shape of the free energy surface was also tested.

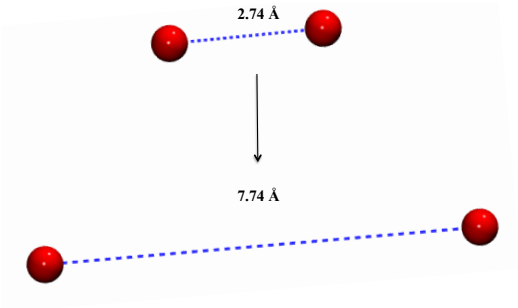

Figure 3.9 A two-particle model system to test the EVB-SBP method

3.5.1 Testing the method

In the following, the effect of using harmonic, Morse and Lennard-Jones potentials to describe native interactions is presented. In the first case, two harmonic potentials (Eq 3.2) were considered, corresponding to the two states with k=0.5 kcal mol^{-1} Å^{-2} and energy minimum r_{ij}^{eq} at 2.74 Å and 7.74 Å. In the second case, a Morse potential (Eq 3.3) was used to describe native interactions where D_e and α were chosen to be D_e=0.5 kcal mol^{-1} and α=0.5 Å^{-1}. Finally, a classical 12-6 Lennard-Jones (Eq. 3.1) potential was used to describe native interactions, with ε=0.5 kcal mol^{-1}. Numerical simulations

consisted of successive energy minimizations of the test system at restrained values of the reaction coordinate until a gradient of 10^{-6} kcal mol^{-1} Å was achieved in steps of r_{ij}=0.05 Å and k=20 kcal mol^{-1} Å^{-2}. The reaction coordinate was chosen to be the inter-particle distance. The EVB parameters used in these calculations were arbitrarily chosen to be $\Delta\alpha_{12}$=0 kcal mol^{-1} and V_{12}=0 kcal mol^{-1}. The diabatic state energies were also calculated using the analytical expressions for harmonic, Morse and Lennard-Jones functions. The resulting double well potentials are shown in Fig. 3.10. The perfect agreement between the data from numerical and analytical calculations indicates that the EVB-SBP method was implemented in `sander` module with correct energy and gradient expressions.

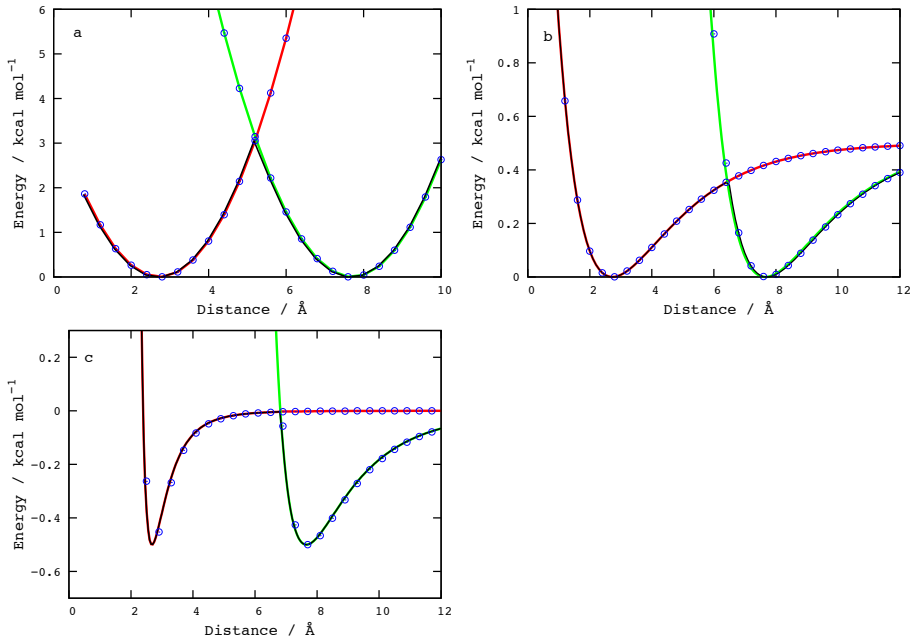

Figure 3.10 Comparison of the results obtained by numerical calculations using the EVB-SBP method and calculated by analytical expressions. Native interactions are modelled by a harmonic potential (a), Morse potential (b) and Lennard Jones potential (c). In each example are shown: diabatic state energies for the initial state (red) and final state (green); EVB ground state energy of the system (black line); results from analytical calculations (blue circles).

The harmonic potentials (Fig. 3.10a) have a minimum at 0 kcal mol^{-1} and increase proportionally to the displacement from the minimum. It should be noted that the harmonic potential has a limitation, since it prevents the system from exploring conformations which are far from the minimum. A more realistic scenario is provided by anharmonic functions such as Morse and Lennard-Jones. Both these potentials are characterized by a stiff short-range repulsive term and a long-range attractive term. The

Morse potential for both minima is equal to 0 kcal mol^{-1} and tends to the separation energy 0.5 kcal mol^{-1} (D_e) at infinite distance. On the other hand, the Lennard Jones potential in both minima is -0.5 kcal mol^{-1} (ε) and tend to 0 kcal mol^{-1} at infinite distances. The Lennard-Jones potential curve is stiffer than the Morse in both the attractive and repulsive terms. Interestingly, the Lennard-Jones potential well is wider at r_{ij}^{eq}=7.74 Å than at 2.74 Å, and hence allows for larger fluctuations at longer distances. Furthermore, due to the large (5 Å) separation of distances between the two minima, larger than those encountered in systems in Chapters 4 and 5, a flat transition region is observed, and an extremely large energy gap range between initial and final states.

In the next section the parameterisation procedure is described using the harmonic potential. This choice is motivated by the perfect symmetry of the potential and the lowest energy gap range between the initial and final states (Δε=25 kcal mol^{-1}) with respect to the other functions (Morse Δε=60 kcal mol^{-1}; Lennard-Jones Δε=117,022 kcal mol^{-1}).

3.5.2 Parameterisation of the potential energy surface

Various potential energy functions (i.e., harmonic, Morse or Lennard Jones) can be used to describe individual native contacts in each state. These functions are then coupled together by EVB to construct a multiple basin potential energy surface. Here a surface is built using two harmonic potentials to investigate the effect of varying specific EVB parameters on the free energy surface, using unbiased and biased simulations. Unrestrained simulations (Fig. 3.11), each 50 ns long, are performed using Langevin dynamics. In addition the system is driven between the two states using umbrella sampling along the energy gap reaction coordinate to calculate free energy of the transition (Fig 3.12). The transition is induced by modifying the energy gap reaction coordinate in steps of Δε=0.02 kcal mol^{-1} and sampling data for 500 ps in the presence of a harmonic restraint k=0.5 mol $kcal^{-1}$. The free energy is then calculated from the instantaneous values of the reaction coordinate and the target value using WHAM (see Chapter 2.). While there is no physical model corresponding to these tests, the main purpose is to elucidate how the diabatic state shift and coupling element can be used to parameterize the free energy surface.

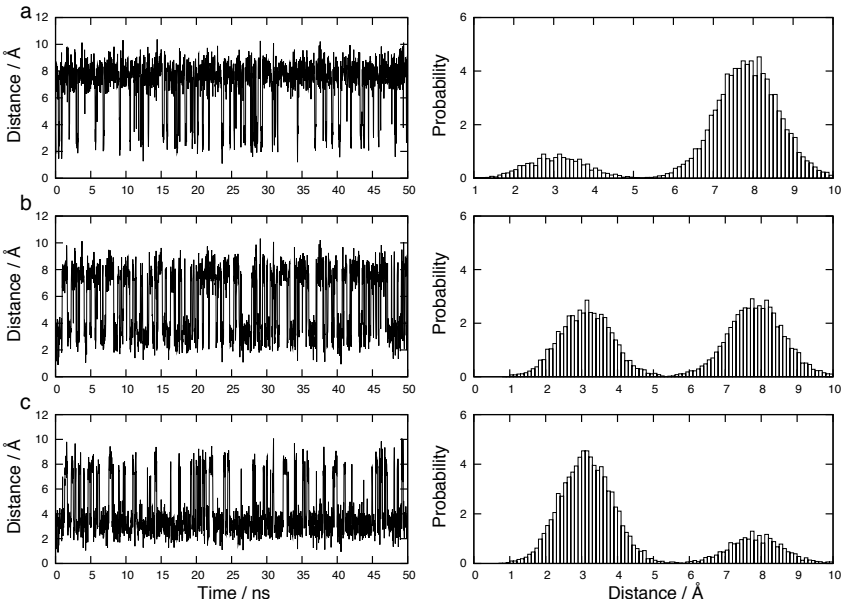

Figure 3.11 Time series of the inter-particle distance (left panel) and the corresponding probability distribution (right panel) are shown for equilibrium simulations with different EVB parameters. (a) $\Delta\alpha$=0 kcal mol^{-1} and V_{12}=0 kcal mol^{-1}; (b) $\Delta\alpha$=1 kcal mol^{-1} and V_{12}=0 kcal mol^{-1}; (c) $\Delta\alpha$=2 kcal mol^{-1} and V_{12}=0 kcal mol^{-1}.

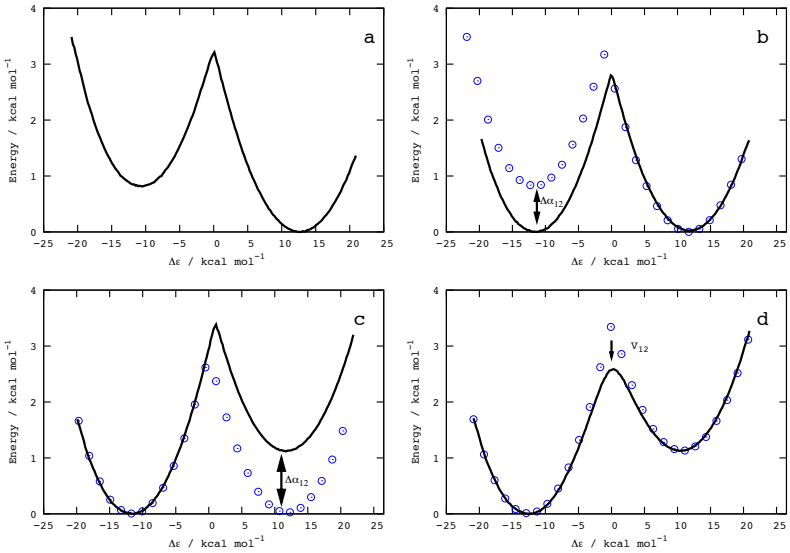

Figure 3.12 The free energy profile is plotted as a function of the energy gap reaction coordinate for four different cases: a) $\Delta\alpha$=0 kcal mol^{-1} and V_{12}=0 kcal mol^{-1}; b) $\Delta\alpha$=1 kcal mol^{-1} and V_{12}=0 kcal mol^{-1}; c) $\Delta\alpha$=2 kcal mol^{-1} and V_{12}=0 kcal mol^{-1}; d) $\Delta\alpha$=2 kcal mol^{-1} and V_{12}=1 kcal mol^{-1}. In graphs (b,c,d), blue circles indicate the free energy profile before the modification of the EVB parameters.

In the first case, both the coupling element and the diabatic state shift are set to zero. Since identical native harmonic potentials are used to define the initial and final states, the ground state energy will be exactly the same in each basin (see Fig. 3.10a above). The equilibrium populations of the two states, however, result in a marked asymmetry at a finite temperature (Fig. 3.11a). The free energy corresponding to the 'final' state,

having longer inter-particle distance, is lower by ~1 kcal mol^{-1} with respect to the 'initial' state (Fig. 3.12a). This can be explained by the fact that thermal motions favour longer inter-particle distances at equal energetic cost; translational motion tends to increase rather than decrease the intervening distance. As noted by Wang et al. [245] the number of conformations accessible for a two-particle system is proportional to r^2 where r is the interparticle distance. The longer interparticle distance has higher number of accessible states, higher probabilities and hence lower free energy. Quantitatively, the observed free energy difference for the two-particle model herein studied can be related to Jacobian factors [246] and their contribution to the configurational partition function. The free energy can be expressed as $\Delta G=-k_B T\ln Z$ where k_B is the Boltzmann constant, T the temperature and Z the partition function. According to Boresch et al. [246], the total configurational free energy is equal to $\Delta G_{TOT}=\Delta G_F+\Delta G_J$, where ΔG_F is the contribution that arises from the force field in a simulation, while ΔG_J is the contribution from equilibrium geometries, or Jacobian factors. The Jacobian factor depends only on geometrical properties (bond length; bond angles) and can be calculated analytically. This allows for the evaluation of the contribution to the total partition function, of what has been defined as "*dynamic stretch free energy*" [247] by simply post-processing simulation results. For a simple two atoms model, the Jacobian factor is $J=V4\pi r^2$ where r is the interparticle distance and V the total volume of the system [246]. Hence the contribution to the free energy can be estimated as:

$$\Delta G_J = -k_B T \ln\left(\frac{J^f}{J^i}\right) \tag{3.6}$$

where ΔG_J is the contribution to the total free energy which arise from the Jacobian factor, k_B the Boltzman constant, T the temperature and J is the Jacobian factor for the initial i and final state f. In the two particle test case herein studied, the contribution to the final conformational free energy has been calculated to be ΔG_J=-1 kcal mol^{-1}, which explains quantitatively the free energy difference observed between the initial and the final state (Fig 3.12a).

Next, the relative energies of the states are changed by varying the diabatic state shift such that the final state is now higher in energy than the initial state by 1 kcal mol^{-1} ($\Delta\alpha$=1 kcal mol^{-1}). In this case, the enthalpic and entropic differences cancel out and the system spends equal amount of time in each state (Fig. 3.11b). Consequently, the initial and final states are calculated to be equally stable in terms of free energy (Fig. 3.12b).

If the diabatic shift is further increased ($\Delta\alpha$=2 kcal mol^{-1}), the final state becomes significantly higher in energy than the initial state and the corresponding population is reduced with respect to the initial state (Fig. 3.11c). The calculated free energy now clearly favours the initial state (Fig. 3.12c).

Finally, the free energy barrier can be modified by introducing a non-zero coupling constant. In this case, the diabatic energy shift was $\Delta\alpha$=2 kcal mol^{-1} and V_{12}=1 kcal mol^{-1}. Note that the introduction of a coupling constant, in the case of a harmonic potential, has no effect on the relative energies of the diabatic states but only on the height of the energy barrier. Thus, this parameter can be applied to reduced the free energy barrier and thus accelerate the kinetics of the transition without affecting the thermodynamic equilibrium of the process (Fig. 3.12d)

This simple test case has shown how the EVB approach can be used to create a unified potential energy surface with more than one minimum, and how the underlying energy surface can be modified using EVB parameters. Both the diabatic state shift and the coupling constant are effective tools, which allow a parameterisation of the free energy surface. It is noted that the free energy obtained in the umbrella sampling simulations above can be mapped on any (geometric) progress variable not just on the (general) energy gap used as the biasing restraint here. Nevertheless, geometric reaction coordinates can also be used directly to drive the conformational transition in the biomolecule on the EVB-SBP potential energy surface (see Chapter 4).

3.6 Improvement of the parameterisation process of the structure-based potential

As discussed above, one of the main difficulties in defining a correct structure-based potential lies in the parameterisation procedure aiming to find the best combination of native interaction list, cut-off and the energy scaling factor. The scaling factor S introduced in the energy function (Eq. 3.1) is crucial to determine the depth of the native LJ interaction between two atoms i and j. In the first approach (Section 3.2) all the native interactions are scaled by a unique S and hence the interactions between the same atom types contribute equally to the stabilisation of the system. An optimal scaling factor was introduced in an attempt to reproduce globally the structural stability and dynamical fluctuations of the entire system.

An alternative approach is herein presented, where the well depth is determined individually for each atom-atom pair interaction, based on the atom-atom distance distribution calculated from a short all-atom force field simulation. As the reference data is obtained from a single simulation of the entire system, it is expected that the combination of all individually parameterised native contacts will correctly reproduce the global structural and dynamical properties of the system. Inter-atomic distances are recorded in the reference simulation and the normalised probability distribution $P(r_{ij})$ will be used to define the depth of the energy well of the inter-atomic potential. This means that if atoms *i* and *j* exhibit low distance fluctuations ('strong' native interaction) then the distribution will be narrow and the resulting interaction potential will be deep. On the other hand large distance fluctuations ('weak' native interaction) will result in a broad distribution and a shallow inter-atomic potential. In the following, two examples are given, illustrating a strong interaction and a weak interaction. These data are derived from the application of this approach to a protein system (see Chapter 6). Two different atom pairs *i-j* were considered, interacting according to the ff99SB Amber force field [248]. The probability distributions of *i-j* distances $P(r_{ij})$ (Fig 3.13a-b) are calculated directly from the simulation, providing the pseudo free energies, $-logP(r_{ij})$ (blue circle in Fig. 3.13c-d).

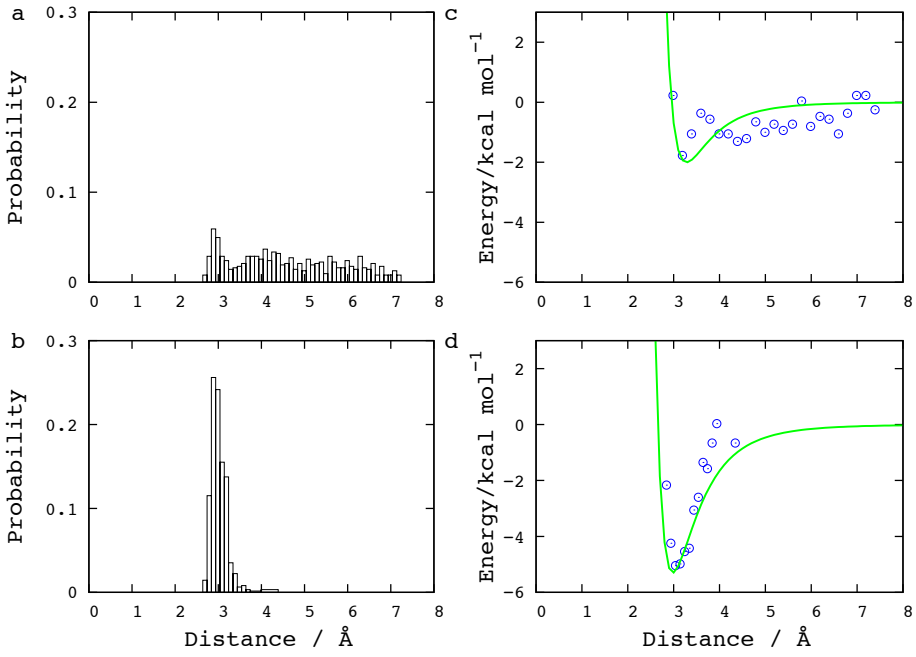

Figure 3.13 The left panels show the probability distribution of the *i-j* inter-atomic distance in two different cases (a,b). The right panels (c,d) show the $-logP(r_{ij})$ values (blue circles) and the fitted LJ potential (green line).

The next step is to introduce a LJ potential which fits the $-logP(r_{ij})$ values. For this purpose, the ε of the Lennard-Jones potential is computed here as the difference between the maximum and the minimum of the $-logP(r_{ij})$ values: $\varepsilon_{ij}=|\log P_{max}-\log P_{min}|$ to obtain the relative free energy for the fluctuation. This will determine the strength of the native interaction defined in the structure-based potential. The result shows that the LJ potential derived with this approach fits reasonably well the $-logP(r_{ij})$ values calculated from the simulation.

The main advantage of this approach is to automate the process of parameterisation for individual atom pairs and incorporate the dynamical effect observed in all-atom force field simulations. This parameterisation method has been successfully applied to reproduce structural and dynamic fluctuations in a protein system (see Chapter 6).

3.7 Summary

In this chapter a novel method developed in this work has been described that utilizes an all-atom structure-based potential (SBP) combined with the empirical valence bond (EVB) theory to study transitions on a multiple-basin energy landscape. This approach can be applied to study conformational transitions where the structural endpoints of the process are known. The first step concerns the parameterisation (cut-off and scaling factor) of the structure-based potentials which are used to describe individual conformations. Then the individual structure-based potentials are combined using the EVB method to build a unified potential energy surface. This requires, two adjustable EVB parameters (adiabatic shift and coupling element) so that the unified potential energy surface reproduces available experimental data. The application of the method to a simple molecular system has allowed for a thorough testing of the implementation of the theory in the *sander* module of the popular Amber 10 simulation package. This development provides the foundation to apply the method to more complex biological systems. In the next chapter, the first application of the method to study base flipping in B-DNA will be described.

Chapter 4

Base flipping in a B-DNA

4.1 Introduction

It is well known that B-DNA is an intrinsically flexible biopolymer, which undergoes local as well as global conformational transitions to perform its biological functions [249]. Structural deformations in DNA are crucial since genetic information is not directly accessible in the duplex state. Base flipping represents one of the simplest structural distortions in DNA and may have different functions [250]. It may represent an initial step in the DNA strands separation and is essential for repair enzymes to have access to bases buried in the DNA helix [251]. In its interaction with proteins, for example, the helical regularity is often disrupted and specific base pairs break up to expose reactive sites of the bases.

Several works, both theoretical and experimental, have been trying to elucidate the structural and energetic insights of the base flipping process. Initially, NMR experiments by Gueron et al. [252] investigated the kinetics of base pair in DNA revealing that base pair lifetimes is in the order of ~10 ms. Further studies also tried to elucidate the mechanism of base flipping upon protein binding [253, 254] to understand whether the protein induces the base flipping or intervene at a later stage on the flipped state. These experiments however provided only a clue to the structural mechanism at the atomic detail. A number of crystal structures of DNA-protein complexes were determined by X-ray [255-257], revealing atomistic information of the structural endpoints of the flipping process. Theoretical studies were also introduced aiming to unravel the structural mechanism at the atomic level. Chen et al. [258] studied the flipping process in DNA with and without the enzyme, suggesting that the preferred pathway was toward the major groove. Subsequently, works from Banavali et al. [85] and Varnai et al. [83] provided more evidence of the flipping mechanism both toward the major and the minor groove. In agreement with these findings, Banavali et al. found

that the opening of a guanine was preferred toward the major groove, while the opening of the corresponding cytosine was symmetric, hence no free energy difference was observed between the minor and major groove pathways. Interestingly, it was also noted the presence of possible intermediates states in the flipping pathway due to the formation of direct or water mediated hydrogen bonds between the flipping base and the neighbour bases [83, 85]. A theoretical study from Hagan et al. [259] also, revealed two main dominant pathways of the flipping mechanism: the first where the fraying of the base pair hydrogen bonds precedes the unstacking of the flipping base; the second where breaking of hydrogen bonds and unstacking happen simultaneously. In addition spontaneous base flipping from equilibrium simulations was also studied using an oligonucleotide containing adenine-difluoro-toluene base pair [84].

One of the best studied examples of base flipping takes place in the B-DNA dodecamer 5'-dGTCAGCGCATGG-3' that contains the target sequence of HhaI DNA C5-methyltransferase [260]. HhaI DNA C5 methyltransferase carries out an extensive base rotation of the central cytosine within a GCGC sequence [261], the mechanism of which was the subject of previous theoretical studies [83, 85, 262]. The first part of the chapter will be focused on the native state dynamics of the canonical B-DNA using structure-based potentials. The parameterisation procedure to reproduce more expensive all-atom force field simulations is discussed in detail. The second part will describe the EVB parameterisation to build a multiple-basin potential and the resulting molecular insight into the base flipping process.

4.2 Parameterisation of the potential based on structural properties

The first aim is to establish the minimal cut-off value that distinguishes between the two conformational states of the system and to keep the size of the native contact list tractable. Once the total number of native contacts is fixed, the scaling factor is varied, with higher value of S resulting in a more attractive native LJ energy term. The value of cut-off has been varied between 3 to 6 Å and the number of native contacts calculated at each cut-off value. The total number of native contacts was divided in three types of interactions: pairing, stacking and cross stacking within the duplex. Given two residues n and m forming a base pair, pairing involves contacts between residues n and m; stacking between n and n+1/n-1, and between m and m+1/m-1; and cross-stacking between residue n and m+1/m-1 or m and n+1/n-1 (numbering from a given base in the

5’-3’ direction within a strand). There were two cases considered in the canonical B-DNA structure: C18-G7 and A16-T9 base pairs (Table 4.1). This structure was built using the nucgen program (part of the Amber package [243]) with standard fibre-diffraction parameters. It should be noted that the exact number of native contacts changes slightly depending on the sequence context of the bases in the target sequence.

Table 4.1. Number of native contacts at different values of cut-off for Watson-Crick base pairs C18-G7 and A16-T9.

Cut-off (Å)	PAIRING	STACKING	CROSS STACKING	TOT
Guanine – Cytosine pair				
3	2	0	0	2
3.5	3	44	4	47
4	15	126	21	162
4.5	17	220	54	291
5	22	336	67	415
5.5	33	444	114	591
6	44	582	161	787
Adenine – Thymine pair				
3	2	0	0	2
3.5	2	45	4	51
4	12	122	16	150
4.5	14	217	45	276
5	19	325	61	405
5.5	27	438	100	565
6	36	569	141	569

A schematic representation of the native interactions involved in base pairing interactions using an atom-atom distance based cut-off of 4.5 Å is shown in Fig. 4.1.

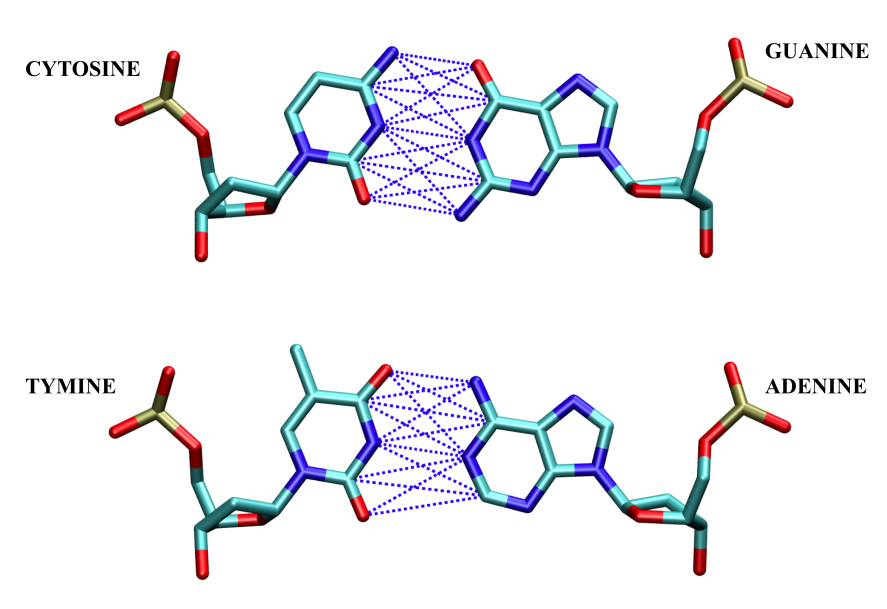

Figure 4.1 Schematic representation of base pairing described by native contacts with a cut-off value of 4.5 Å between atoms of C-G and T-A base pairs.

A large number of native contacts are visible between atoms on the Watson-Crick edge of the bases, with a GC pair having 17 contacts while an AT pair having 14 contacts, consistent with the larger number of hydrogen bonds in GC pairs. A structural representation of all native interactions involved in pairing, stacking and cross-stacking is shown in Fig. 4.2. It is clear that the major contribution to the total number of contacts stems from the native stacking interactions. A difference in the cross-stacking contacts between purines (adenine, guanine) and pyrimidines (thymine, cytosine) is visible (Fig. 4.2). The imidazole ring of purines clearly enhances cross-stacking interactions compared with pyrimidines. The larger number of native contacts in stacking rather than pairing bases is consistent with the fact that stacking of bases is preferred over pairing in aqueous solution [263]. By summing up the native contacts for a given base pair it appears that a simple atomistic structure-based potential is capable of reproducing the energetic trend between GC and AT base pairs with N=291 and 276, respectively.

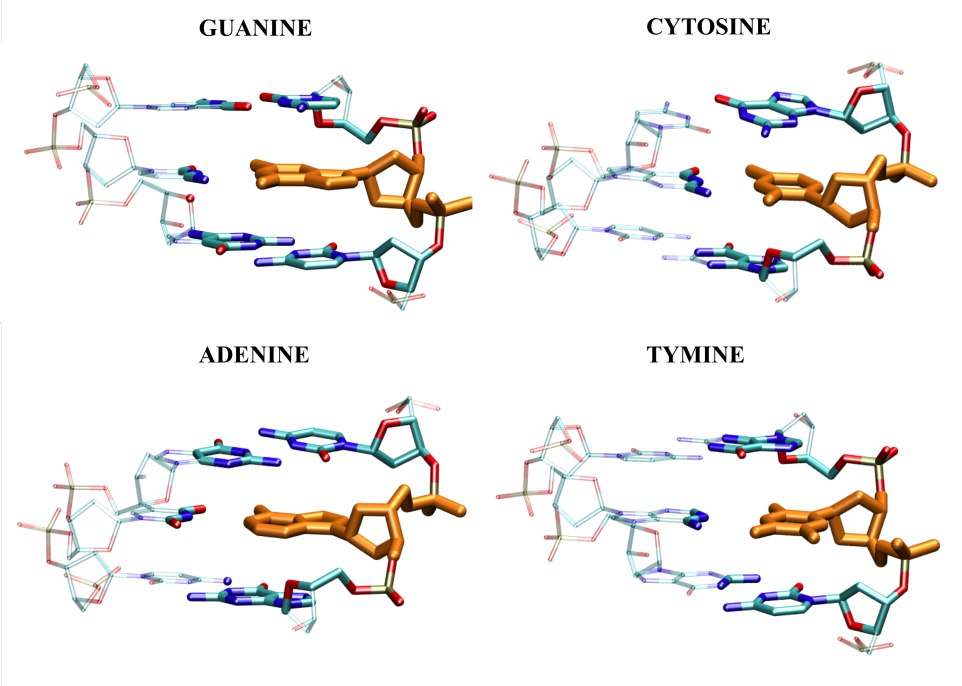

Figure 4.2. Schematic representation of the heavy atoms within 4.5 Å from a given base (guanine, cytosine, adenine and thymine in orange) is shown in atom-coloured opaque representation.

The actual form of the native interaction used in this study is a classic 12-6 Lennard-Jones potential. In principle, the native energy should replace the full non-bonded energy term of a classical force field that includes intra-molecular electrostatic interactions and solvation effects as well. The expectation here is that a sum of scaled Lennard-Jones potentials allows the recovery of the full non-bonded term of the force field.

The parameterisation procedure involves performing a series of short (5 ns) Langevin dynamics simulations of the B-DNA at 300 K and low friction (γ=5 ps^{-1}) using the EVB-SBP method. In these simulations, the native contact list is determined by a cut-off value in the range between 3.5 Å and 6 Å; the equilibrium distances are given by the sum of the van der Waals radii of the atoms in contact from the parm99.dat parameter file of the Amber force field, and the scaling factor is varied in the range between 1 and 5. The average and standard deviation of the root mean square deviation (RMSD) of the resulting structural ensemble are then calculated for each value of *S* and compared with that of the classical force field results (Fig. 4.3). Newtonian dynamics simulation of the B-DNA in explicit water and counterion environment with the Amber parm99 [153] force field was previously performed at constant temperature (300 K) and

pressure (1 bar) [264]. It has to be noted that a more recent version of the Amber force field, parmbsc0 [152] has been introduced to correct the overpopulation of α/γ=g+/t backbone angles observed using parm99. However, this effect was observed for simulations longer than 10 ns, and the reference simulation herein employed is only 5 ns. The RMSD represent a simple measure to assess overall structural stability in the test simulations with respect to the reference model.

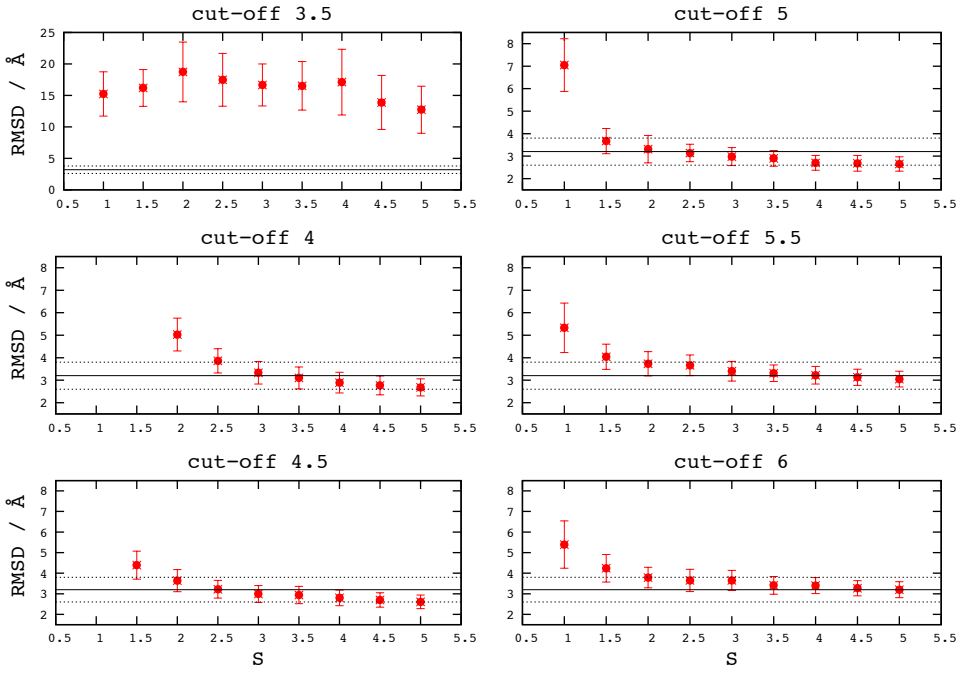

Figure 4.3 Calculated average root mean square deviation (RMSD) of the Cartesian coordinates of all atoms in the DNA duplex from the canonical B-DNA simulated using the structure-based potential with different cut-off and scaling factors, *S*; error bars indicate one standard deviation from the average. For reference, the result using the parm99 force field [153] is shown as a horizontal solid line with dashed lines indicating one standard deviation from the average. Note that only for cut-off 3.5 Å the y range is increased to 25 Å due to a completely unfolded structure at each *S* value. In addition, in the remaining graphs, points for *S* values with high average RMSD (>14 Å) are excluded. The points discarded are: cut-off 4 Å (*S*=1;*S*=1.5); cut-off 4.5 Å (*S*=1).

In general, if the native contact energy contribution is too low (low *S* values and/or low cut-off) then partial or total unfolding of the structure occurs. On the other hand, if the native contact energy contribution is too high (high *S* values and/or high cut-off values), the structure remains very close (in terms of RMSD) with respect to the reference structure, although fluctuation will be unreasonably small, resulting in "frozen" structures (Fig. 4.4).

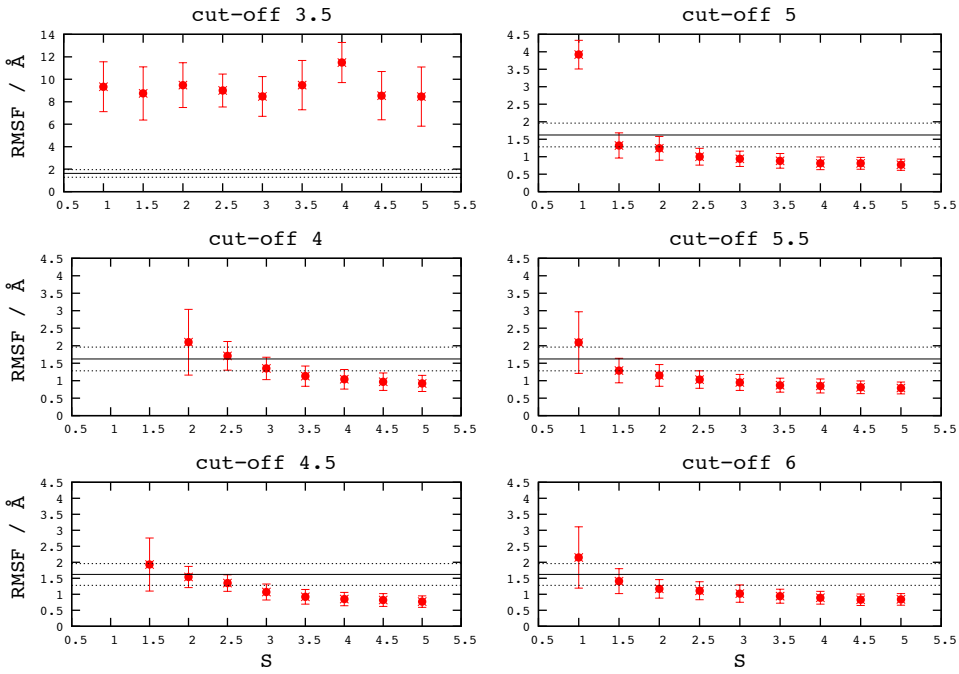

Figure 4.4 Calculated average per residue root mean square fluctuation (RMSF) from the canonical B-DNA simulated using the structure-based potential with different cut-off and scaling factors, *S*; error bars indicate one standard deviation from the average. For reference, the result using the parm99 force field [152] is shown as a horizontal solid line with dashed lines indicating one standard deviation from the average. Note that only for cut-off 3.5 Å the y range is increased to 14 Å due to high (>6 Å) fluctuations at each *S* value. In addition, in the remaining graphs, points for *S* values having high RMSF (>4.5 Å) are excluded in order to increase resolution. The points discarded are: cut-off 4 Å (*S*=1;*S*=1.5); cut-off 4.5 Å (*S*=1).

It is found that the best compromise in terms of both RMSD and RMSF is cut-off 4.5 Å and rescaling factor *S*=2.5. These parameters reproduce the force field results remarkably well and will be used to study the native state dynamics and base flipping in B-DNA. In summary, a structure-based potential with a relatively short cut-off value for native interactions appears to reproduce results from more accurate and expensive force field simulations.

4.2.1 Parameterisation of the potential based on energetic properties

An alternative way to parameterise the structure-based potential (SBP) would be to reproduce the relative energies of pairs of molecular conformations and thus the energy gradient from force field (FF) simulations. Let us briefly recall the main differences between the FF and SBP energy functions. The bonded energy term (bond, angle, and dihedral) is identical in both energy functions, while the non-bonded part is different. In

the FF, the latter is the sum of electrostatic, van der Waals and solvation energies, while in the SBP it is the sum of the non-native repulsion and the native attraction energies (see Chapter 3). Consequently, the aim here is to find a combination of cut-off and scaling factor values for the SBP that compensates the non-bonded FF energy terms. Although the cut-off value influences slightly the non-native energy contribution, the key contribution stems from tuning the native interaction energy. Thus the premise of this work is that differences in a scaled Lennard-Jones energy for native interactions can reproduce the non-bonded energy differences of a FF between pairs of structures. A brief description of the approach used here and the results obtained are presented in the following.

Two structural ensembles were generated: 500 structures from a 50 ns FF simulation (see details above) provided a "folded ensemble" and 500 structures from an unfolded simulation with the DNA helix completely unfolded provided an "unfolded ensemble". This latter was generated by performing SBP simulation using a cut-off=4.5 Å and a scaling factor, *S*=1. Structures within the folded ensemble resemble strongly but the associated energy differences showed large fluctuations rendering the analysis pront to error. Therefore, the energy difference between each structure of the folded ensemble with each structure of the unfolded ensemble was evaluated. The energy was calculated using two approaches. In the first case, the standard MM-PBSA method [243] was employed, which combines the internal molecular mechanical (MM) energy calculated from the FF with Poisson-Boltzmann solvation (PB). The latter term includes both the electrostatic contribution calculated using a numerical solver for the Poisson-Boltzmann equations [173] and the non-polar contribution calculated using a solvent-accessible surface area (SA) dependent term [265]. In the second case, the energy was evaluated using the SBP with different combinations of cut-off and scaling factors: cut-off range between 3.5-6 Å and scaling factors between *S*=0.5-5.

Thus, the aim was to find the parameters of the SBP that reproduce the force field energy difference and hence the gradient between any structures of the two ensembles. In each case, the result was quantified by a quality factor, Q [266]:

$$Q = \frac{\sqrt{\sum (\Delta E_{SBP} - \Delta E_{FF})^2}}{\sqrt{\sum (\Delta E_{FF})^2}} \tag{4.1}$$

where ΔE_{SBP} is the energy difference between a given folded and unfolded structure calculated using the SBP and ΔE_{FF} the reference energy difference calculated using MM-PBSA. A low Q-factor (~0) indicates high degree of similarity between the two ensembles.

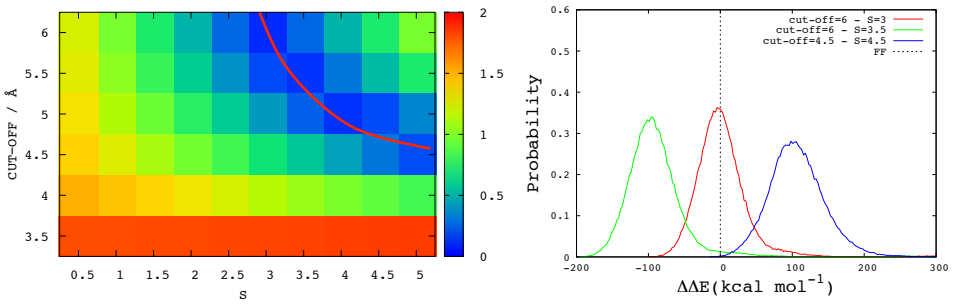

Figure 4.5 A comparison between the energy differences evaluated by the MM-PBSA and SBP methods. (left panel) Pictorial representation of Q-factors for different combinations of cut-off and scaling factor (S) parameters to evaluate ΔE_{SBP}. Red line has been drawn to highlight the region with low Q-factors. (right panel) The $\Delta\Delta E$ probability distribution is shown for three different cases.

The analysis reveals a region with low Q-factors of <0.5 (Fig. 4.5) when scaling factor is between 3 to 5 and cut-off 4.5 Å to 6 Å. The corresponding favourable structural properties (RMSD and RMSF) are shown above (Figs. 4.3, 4.4). Both structural and energetic parameterisation evidence that a cut-off 3.5 Å in SBP is too short to include essential interactions of the molecule, resulting in unfolded structures and poor similarity in energy differences (Q-factor ~2). A scaling factor of S=1 and lower values are also inconsistent with FF results (Q-factor >1.5). This is reasonable considering the fact that the stabilising electrostatic and solvation terms in the FF can only be reproduced with increased attractive interactions in the SBP. Interestingly, some combinations of cut-off and S parameters result in an intermediate Q-factor (~1), but good structural correspondence compared to the FF (see, for example, cut-off=4 Å / S=3; 4.5 Å / 2.5; and 5 Å / 2).

In order to depict a more quantitative picture of the difference between ΔE_{SBP} and ΔE_{FF} the $\Delta\Delta E$ is calculated as $\Delta\Delta E = \Delta E_{SBP} - \Delta E_{FF}$. If the two values are similar, $\Delta E_{SBP} \sim \Delta E_{FF}$, then $\Delta\Delta E$ is distributed closely around 0 kcal mol^{-1}. The best agreement is obtained with parameters cut-off=6 Å and S=3 (Q-factor=0.1) indicating similar gradients between FF and SBP (right panel, Fig. 4.5). Alternatively, in case $|\Delta E_{SBP}| < |\Delta E_{FF}|$, the native energy is not sufficient to compensate for the FF non-bonded

energy; see, for example, cut-off=4.5 Å and *S*=4.5 (Q-factor=0.3), the $\Delta\Delta E$ distribution is centred at +100 kcal mol^{-1}. If $|\Delta E_{SBP}| > |\Delta E_{FF}|$, the native energy is overcompensating the FF non-bonded energy; cut-off=6 Å and *S*=3.5 (Q-factor=0.26), the $\Delta\Delta E$ distribution is centred at -100 kcal mol^{-1}.

Table 4.2 Non-bonded energy components calculated using FF and SBP (cut-off=6 Å and *S*=3) for six selected pairs of structures. Energy values are reported in kcal mol^{-1}.

		ΔE_{nn}[1]	ΔE_{n}	ΔE_{ele}	$\Delta E_{vdw1\text{-}4}$[2]	$\Delta E_{ele1\text{-}4}$	ΔE_{PB}	ΔE_{tot}	$\Delta\Delta E$
1	**FF**	-178.19		-409.42	50.83	-76.43	232.54	-380.67	92.67
	SBP	90.17	-456.52	0	77.91	0	0	-288.44	
2	**FF**	-177		10.86	38.65	-105.8	-122.2	-355.52	114.51
	SBP	97.60	-416.52	0	59.02	0	0	-241.01	
3	**FF**	-225.79		-415.21	44.69	-40.26	219.79	-416.78	0.67
	SBP	81.53	-564.36	0	66.72	0	0	-416.11	
4	**FF**	-240.46		249.87	55.36	-79.87	-379.4	-394.45	0.4
	SBP	98.59	-575.16	0	82.51	0	0	-394.05	
5	**FF**	-243.68		486.05	46.05	-55.27	-582.1	-348.91	-101.8
	SBP	85.51	-605.04	0	68.82	0	0	-450.71	
6	**FF**	-249.33		425.61	42.80	-44.04	-519.9	-344.77	-74.09
	SBP	97.17	-592.44	0	66.41	0	0	-428.86	

[1] ΔE_{nn} and ΔE_{n} represent the native and non-native interactions in SBP. Single values for FF show the van der Waals energy.
[2] $\Delta E_{vdw\ 1\text{-}4}$ in the SBP include only the repulsive term and in the FF both attractive and repulsive.
[3] ΔE_{PB} include both polar and non-polar terms to the Poisson-Boltzmann solvation free energy.

To understand how the difference between the two energy models arises, the energy components of ΔE in the FF and SBP are compared (using cut-off=6 Å and *S*=3) for individual pairs of structures (Table 4.2). It is notable that even with an optimal parameter set and $\Delta\Delta E$ ~0 kcal mol^{-1}, the distribution is rather wide and one can observe the three scenarios discussed above.

In the first two cases (Table 4.2) the native energy in SBP is not sufficient to compensate for the non-bonded energy in the FF; in cases 3 and 4, the ΔE_{SBP} and ΔE_{FF} are almost identical; and in cases 5 and 6 the native energy in SBP is more attractive compared to the non-bonded energy of the FF. This example shows the limitations of the current SBP using a single, scaled Lennard-Jones term for native interactions to reproduce the energetic properties of the more accurate FF.

In summary, according to the energetic analysis the parameters cut-off=6 Å and *S*=3 appear to be optimal to reproduce the energy difference between folded and unfolded ensembles. However, these parameters showed slightly higher RMSD and reduced fluctuations compared to the FF for the folded ensemble (see above). In contrast, the parameters cut-off=4.5 Å and *S*=2.5 appeared to be optimal in terms of structural data for the folded ensemble. A SBP is primarily employed in this thesis to reproduce a structural ensemble of a reference FF and not to describe structural transitions between different ensembles. The relative energy of the structural ensembles is described by the EVB parameters (see below). Since a lower cut-off represents lower computational cost, the parameters, cut-off=4.5 Å and *S*=2.5 were considered to be a reasonable balance between structural, dynamical and energetic properties of the system.

4.3 Native state dynamics of a B-DNA

We illustrate how molecular dynamics simulations using a simple structure-based potential (SBP) are capable of reproducing the reference structural ensemble of a B-DNA generated by an equilibrium simulation [264] using an all-atom force field for nucleic acids [153]. The root mean square deviation (RMSD) of the Cartesian coordinates of all atoms in the DNA duplex from the canonical B-DNA simulated using the structure-based potential (SBP) or the force field (FF) is practically indistinguishable with an average value of 3.2±0.6 Å and 3.4±0.5 Å for FF and SBP, respectively (Fig. 4.6). Positional fluctuations of residues from the respective average structure (RMSF) along the sequence were slightly but consistently lower for SBP with respect to FF data (Fig. 4.6).

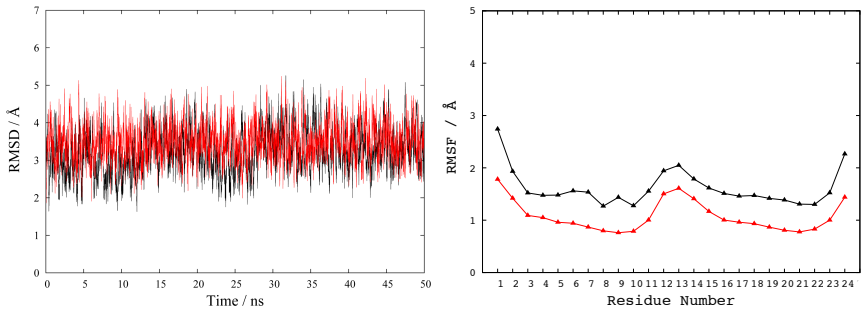

Figure 4.6 Root mean square deviation of atomic positions from the canonical B-DNA (left panel); and root mean square fluctuations of residues around the average structure (right panel) calculated from 50 ns simulations using structure-based potential (red line) and force field (black line).

Inter-base pair parameters, which describe rotations and translations between successive base pairs along the x-axis (tilt, shift), y-axis (roll, slide) and the z-axis (twist, rise) have been calculated with respect to an optimal global helix axis, using the CURVES program [267]. Output values are compared in Fig. 4.7. It is clear that not only the average values of the helical parameters calculated using SBP reproduce those from FF simulations and experimental data (Table 4.3) but even their distributions are well matched (Fig. 4.7). Based on these results it is concluded that a simple structure-based potential can describe structural and dynamical properties of B-DNA in detail.

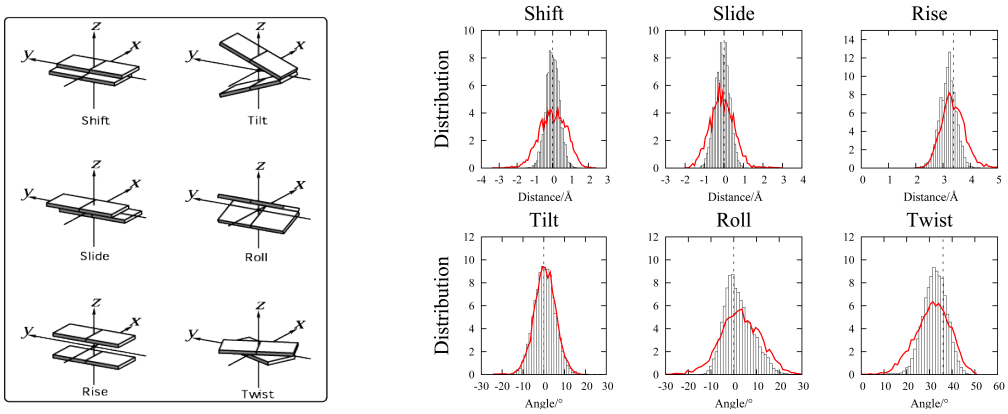

Figure 4.7 Left panel shows schematic representation of rotations and translations between successive base pairs along the x-axis (tilt, shift), y-axis (roll, slide) and the z-axis (twist, rise) [268]. Right panel shows histograms of inter-base pair parameters calculated using structure-based potential (black bin) and force field simulation (red line). The dashed line represents the canonical value for B-DNA [269].

Table 4.3. Average helical parameters and standard deviations of a B-DNA double helix calculated from 50 ns simulated trajectories using force field (FF) and structure-based potential (SBP), compared with experimental data [269]. Translations are shown in Å and rotations in degrees.

	FF	SBP	Exp
Shift	0.0 ±0.7	0.0±0.4	-0.02±0.45
Slide	-0.1 ±0.6	-0.1 ±0.4	0.23±0.81
Rise	3.4 ±0.4	3.2 ±0.3	3.32±0.19
Tilt	0.6±5.2	0.3±5.2	-0.1±2.5
Roll	3.0±8.9	2.2±6.4	0.6±5.2
Twist	31.2±8.0	32.4±5.1	36.5±6.6

4.4 Parameterisation of the two-basins potential

Once the single basin has been parameterized the next step to simulate base flipping in B-DNA is to construct a two-basin potential that corresponds to the endpoints of the base flipping process: the fully stacked and the flipped state. Each state is defined by a unique structure-based potential. The reference structure used to define the potential was a canonical B-DNA for the stacked state (Fig. 4.8), while DNA in the crystal structure of the protein-DNA complex (pdb id 6mht) was used for the flipped state. In this crystal structure the central cytosine (C18) is completely flipped out of the helical stack (Fig. 4.8) [255].

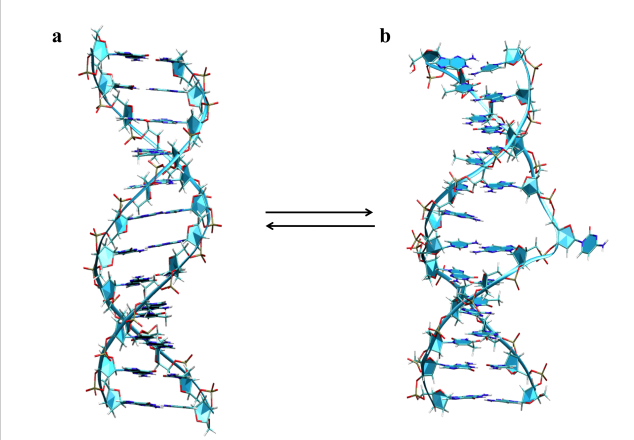

Figure 4.8 Schematic representation of the two endpoints of the base flipping process: B-DNA in the closed state (a) and in the flipped state (b).

Using a cut-off value of 4.5 Å, a significantly higher number of native contacts is obtained for the closed form (N=1665) than for the flipped form (N=1241). Since the stability of each structure is directly related to the total number of such contacts, the closed state is significantly lower in energy compared to the flipped state.

As shown in the previous chapter (Chapter 3), EVB parameters can be used to alter the shape of the potential energy landscape to reproduce available experimental or theoretical data. Following the two particles test case, herein the same procedure has been applied to the base flipping which represent a more complex process.

To illustrate the ease of use of EVB parameters in constructing the energy landscape of interest, in Fig. 4.9 is shown the effect of varying the energy shift, $\Delta\alpha$, and the constant coupling element, V_{12}. It is clear that while shift changes predominantly the energy difference between the initial and final states, the coupling element primarily

changes the transition state region. It should, however, be noted that by changing these parameters, we effectively modify the entire energy landscape. For example, a change in V_{12} also changes the position of the maximum of the energy profile. Furthermore, structure-based potentials are highly non-harmonic functions of the system coordinates, and thus the transition state does not need to be at zero reaction coordinate, in contrast with electron or proton transfer processes with mostly harmonic fluctuations [270].

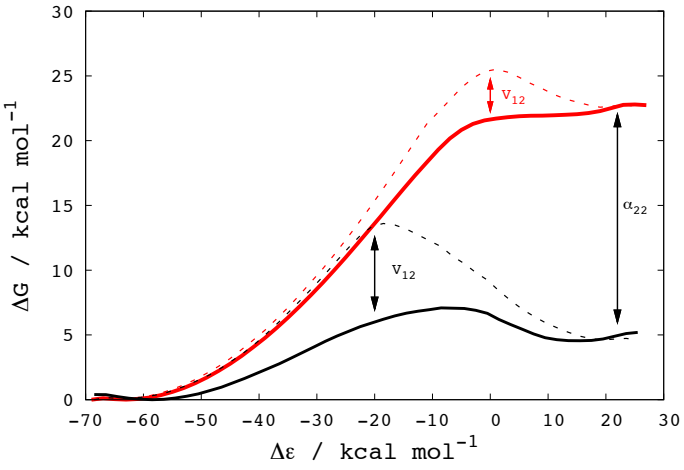

Figure 4.9 Effect of different simulation parameters in the EVB-SBP method to alter the free energy surface of conformational transitions: 1) $\Delta\alpha_{12}$=-68 kcal mol^{-1} and V_{12}=2 kcal mol^{-1} in black dashed line; 2) $\Delta\alpha_{12}$=-68 kcal mol^{-1} and V_{12}=13 kcal mol^{-1} in black solid line; 3) $\Delta\alpha_{12}$=-48 kcal mol^{-1} and V_{12}=2 kcal mol^{-1} in red dashed line; 4) $\Delta\alpha_{12}$=-48 kcal mol^{-1} and V_{12}=8 kcal mol^{-1} in red solid line.

The easy construction of energy landscapes can be particularly useful as a simple means to test the effect of multiple sequence mutation on the mechanistic aspects of the transition. Similarly, by reducing the kinetic barrier to structural changes, the frequency of rare events can be increased and hence direct simulation of large-scale conformational changes becomes possible.

4.5 Base flipping in B-DNA: insight into the mechanism

For the base flipping process, energetic data from earlier free energy calculations performed on the same system using all-atom force field and explicit environment [83], are used as reference for the EVB parameterisation procedure. It is however emphasised that, in general, energetic data should be obtained from experimental kinetic and thermodynamic data and hence there is no need to carry out simulations with force field prior to using the EVB-SBP method. It was established that the flipped state does not

correspond to a stable high-energy structure but rather to a plateau at ~22 kcal mol^{-1} above the closed state. The lack of a significant barrier to base closing is consistent with the experimental observation that open-state lifetime is on the nanosecond timescale [82, 271]. It should, however, be noted that DNA "breathing" or spontaneous base opening that can be detected by imino proton exchange in NMR spectroscopy corresponds to a more limited base rotation out of the helical stack [272-274]. While spontaneous base opening is a process on the millisecond timescale [82], the kinetics of a non-enzymatic base flipping is several orders of magnitude slower, as evidenced by DNA denaturation in the presence of β-cyclodextrin [275].

To create such a free energy landscape, an energy shift of $\Delta\alpha_{12}$=-48 kcal mol^{-1} and a constant coupling element of V_{12}=8 kcal mol^{-1} has been introduced which reproduces the correct free energy change associated with the process to smooth the surface at the transition state region (Fig. 4.10).

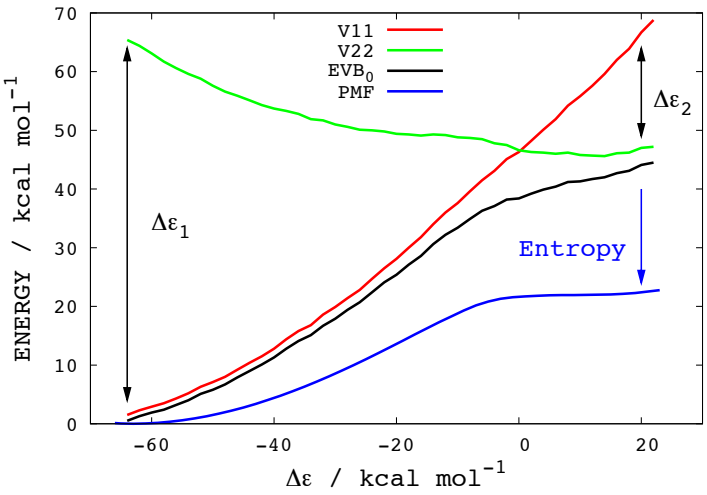

Figure 4.10 Diabatic state energies (V11,V22) and the adiabatic energy (EVB$_0$) are shown along the reaction coordinate. The results represent an average over an ensemble of 8000 structures at a given value of the reaction coordinate. The reaction coordinate is the energy gap (Δε) that is the difference between the potential energies of the diabatic states at a given structure, and range between Δε$_1$ (closed state) and Δε$_2$ (flipped state). The average potential of mean force (PMF) is compared with average ground state energy.

In total, 20 independent simulations in both the forward and reverse directions of the base flipping process have been carried out, by driving the system along the energy gap reaction coordinate. The base flipping or closing was induced by modifying the energy gap reaction coordinate in 2 kcal mol^{-1} steps and continued sampling to collect the instantaneous values of the reaction coordinate for 2 ns. Each simulation consisted

of 90 ns dynamics, totalling 1.8 μs for the whole simulation ensemble. The diabatic state energies, the ground state energy and free energy averaged over the multiple trajectories are shown in Fig. 4.10. The ground state energy and free energy of the process are aligned in correspondence of the closed state where the entropic effect is expected to be minimal compared to the flipped state. In the flipped state the C18 is free to rotate and explore a much larger conformational space than in the stacked state. This represents a stabilizing effect of the flipped state in the energetically unfavorable flipped state. Biased simulations to enhance the sampling along the structural transition were performed using the umbrella sampling technique [210].

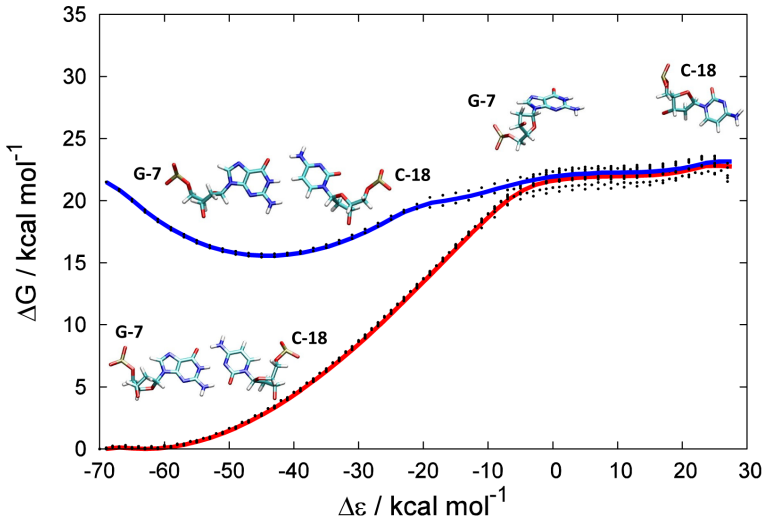

Figure 4.11 Free energy changes of base flipping in B-DNA calculated using the EVB-SBP method. Dotted line shows results from individual molecular trajectories; solid line represents the ensemble-averaged free energy profiles. Opening/closing pathway with the conventional *anti* C18 is depicted in red; pathway with *syn* C18 is shown in blue. The G:C base-pair undergoing opening is shown at the relevant location of the free energy diagram.

The individual trajectories all result in free energy profiles that fluctuate within 1-2 kcal mol^{-1} with respect to the average profile, indicative of well-converged simulations on a smooth energy surface (Fig. 4.11). The ensemble-averaged free energy profile obtained using EVB-SBP reproduce the reference data [83] from FF simulations well. Experimental studies can only indirectly assess the base flipping pathway. Based on structural and kinetic studies, flipping via the minor groove was predicted in the enzyme [260, 276]. Early modelling studies [277-279] of B-DNA suggested that base opens in the major groove direction, which was supported by several simulation studies where spontaneous base opening via the major groove was observed [84, 259, 280]. A

previous study [83, 281], however, using atomistic force field and explicit water environment showed that both major and minor groove pathways are feasible. In that study a biasing potential is used to directly control the opening direction into either the major or minor groove. The authors [83] suggest that although flipping through the minor groove causes some steric clashes of the exocyclic groups, it effectively shields bases from solvent exposure. The energy gap reaction coordinate applied in the present study does not bias the flipping directionality in any way. Remarkably, umbrella sampling simulations of opening and closing processes reveal pathways corresponding to base rotation via both minor and major groove. The direction of the flipping can be followed along the opening angle [83, 282] defined as the angle between the glycosidic bond of the flipping base and the C1'-C1' axis of the corresponding base pair projected to a plane perpendicular to the local helical axis. The opening angle in the closed state is approximately ~50°. The opening toward the major groove results in positive opening angle values, while negative values are observed during the opening toward the minor groove (Fig. 4.12).

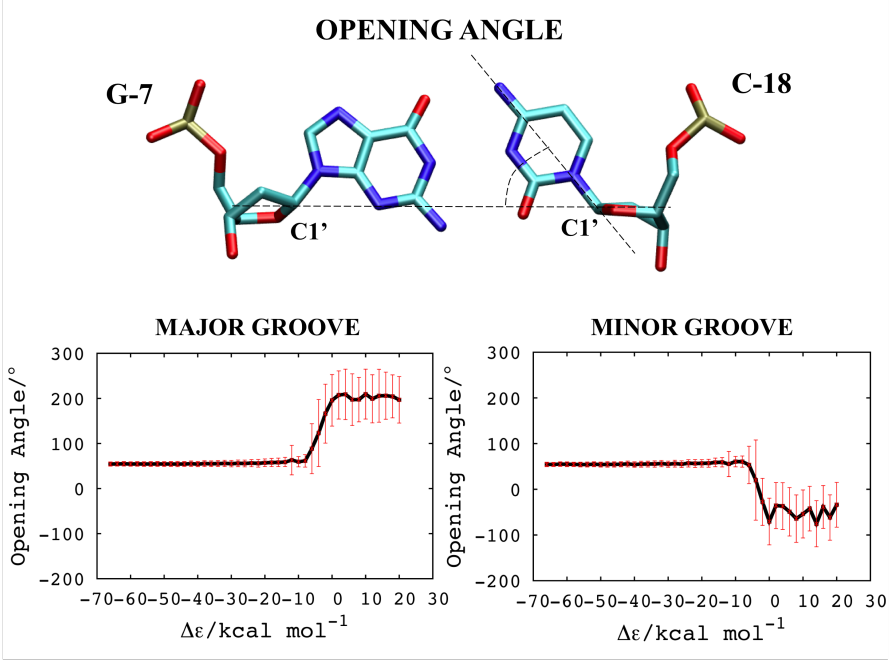

Figure 4.12 Opening angle of C18 toward the major and the minor grooves is shown.

In agreement with previous findings [85], no free energy difference is observed between the opening toward the minor and major groove. However, the major groove pathway appears to be the preferred direction with only 25% of trajectories going

through the minor groove (Fig 4.13). This would confirm previous studies [84, 259, 280, 283] were the opening toward the minor groove was shown to be less favorable due to the presence of a steric barrier, which requires additional backbone deformation in order for the minor groove route to take place.

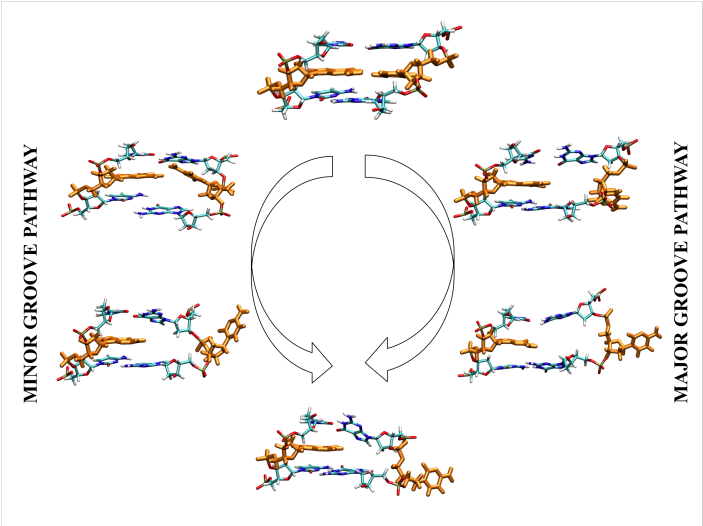

Figure 4.13 Structural snapshots of C18 flipping toward minor and major groove direction.

Interestingly, an alternative closing pathway of the flipped C18 has also been observed, leading to a different conformation of the cytosine (Fig. 4.14). Free energy calculations indicated a local minimum that corresponds to a base stacked in the helical duplex but at ~15 kcal mol^{-1} above the canonical B-DNA state. A careful analysis of the structural properties revealed a closing pathway with C18 in syn conformation as evidenced by the chi angle of 60° (Fig 4.14).

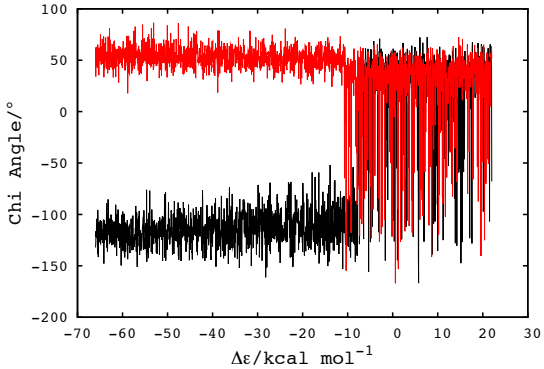

Figure 4.14 The chi angle is plotted along the energy gap reaction coordinate for two different closing trajectories: in red is depicted a trajectory ending in *syn* conformation and in black a trajectory ending in anti conformation. The closed (final) state is at Δε=-60 kcal mol^{-1} and the open (starting) state is at Δε=20 kcal mol^{-1}.

In the flipped state, the base can freely rotate around the glycosidic C1'-N1 bond and thus a closing process may result in a 'dead-end' if initiated from a syn nucleotide. This off-pathway intermediate must re-open before it can re-close correctly. The higher free energy of the duplex with a closed C18(syn)-G7(anti) base pair is due to the altered orientation of C18, which allows for stacking to take place but not proper Watson-Crick hydrogen bonding (Fig. 4.15). The rotation around the glycosidic bond results in the syn conformation, with hydrogen bond between O6 and H2-N4 (Fig. 4.15), while no hydrogen bond is formed between N1-H1 and N3 and between N2-H2 and O2.

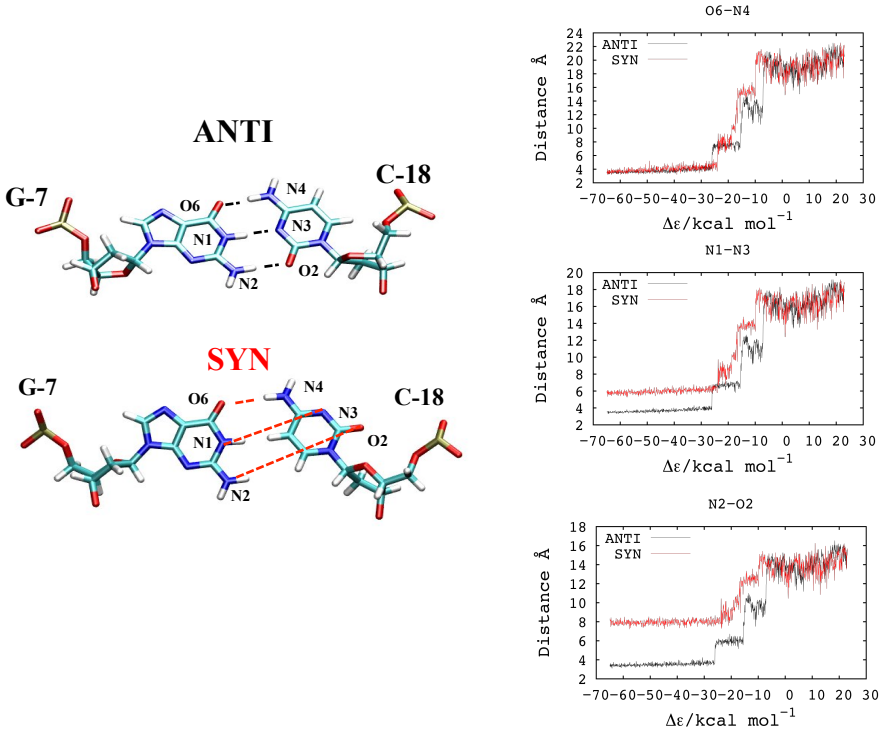

Figure 4.15 Comparison of closing trajectories ending in anti or syn conformation of C18 is shown. The graphs on the right show the distance between the heavy atoms involved in the hydrogen bonding between G7 and C18 (O6-N4; N1-N3; N2-O2) as a function of the energy gap reaction coordinate, for trajectories ending in *anti* (black) and *syn* (red) conformations. On the left the atomic detail of each conformation is shown.

Importantly, it is noted that the free energy obtained in the umbrella sampling simulations above can be mapped on any (geometric) progress variable not just on the (general) energy gap used as the biasing restraint. Nevertheless, geometric reaction coordinates can also be used to directly drive the conformational transition in the biomolecule on the EVB-SBP potential energy surface. To illustrate this point, the free energy is calculated along a new reaction coordinate using structures obtained from

sampling along the energy gap coordinate (Fig. 4.16). This involves calculating the root mean square distance (after least squares fitting) between the actual atomic positions and the crystal structure of the flipped state ("target"); here the heavy atoms of the flipping base and its 3'/5' neighbours are used to define the RMSD reaction coordinate. The free energy profile mapped on the RMSD reaction coordinate are compared with those that are calculated when the RMSD restraint is used directly to drive the base flipping (targeted MD) on the EVB-SBP or force field [281] potential energy surface.

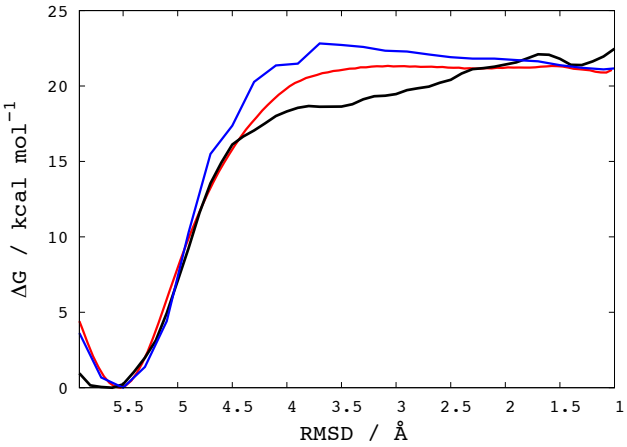

Figure 4.16 Free energy changes of base flipping in B-DNA calculated along an RMSD reaction coordinate with different approaches. Free energy calculated from simulations biased by the energy gap reaction coordinate on the EVB-SBP surface (blue); free energy calculated from simulations biased directly by RMSD on the EVB-SBP surface (red) and on a surface defined by force field [281] (black).

Although these free energy profiles are obtained with significantly different approaches, they all show a minimum at RMSD=5.5 Å corresponding to the closed B-DNA that monotonically increases to about 22 kcal mol^{-1} for the flipped state. These results demonstrate that the EVB-SBP surface matches the potential energy surface defined by the force field [281] reasonably well along different reaction coordinates and that sampling along the energy gap coordinate generates structural ensembles that can subsequently be used to re-map the free energy along any chosen reaction coordinate efficiently.

4.6 Computational efficiency

One of the main advantages of the EVB-SBP approach, compared to conventional all-atom force fields, lies in its significantly reduced computational requirement. This

method was benchmarked using the modified sander module of Amber v10 on four Intel Q6600 CPUs. A standard 1 ns molecular dynamics simulation of the DNA dodecamer with ~5800 explicit water and 22 Na+ in NPT ensemble using the highly optimised `pmemd` module of Amber v10 requires 490 min. As a comparison, the EVB-SBP method took only 18 min to complete 1 ns, a speed up of ~27 times. Consequently, this method can be used to simulate at atomic detail much larger biological systems with significantly improved computational efficiency.

4.6 Summary

The first application of the novel EVB-SBP method has been described to base flipping in B-DNA, an essential but localized conformational change. The method has been shown to be computationally efficient yet fairly accurate to unravel the base flipping mechanism in B-DNA. A procedure has been outlined how to determine cut-off and rescaling factor in order to define a system-specific structure-based potential, which reproduces the structure and dynamical fluctuations of a stable structure from more expensive all-atom force field simulations. Structure-based potentials have been coupled by a parameterised EVB procedure resulting in a free energy landscape that reproduces the salient features of reference systems. It is noted that due to the inherent facility to define energy landscape using this method, it may be used to test mechanistic changes expected due to sequence mutation in the future. The method has been applied in conjunction with umbrella sampling Langevin dynamic simulations using the general "energy gap" reaction coordinate that implicitly includes all the degrees of freedom in the system. This eliminates the cumbersome requirement to define specific geometric progress variable to drive the conformational change *a priori*. Several independent umbrella simulations showed rapid convergence of the free energy that reproduces the reference data. Base rotation was observed via both grooves of the B-DNA duplex with a preference for the major groove pathway. An alternative closing pathway to a high-energy off-pathway intermediate has also been identified that may appear if the process is initiated from a flipped syn base.

Chapter 5

Switching mechanism of a bistable RNA

5.1 Introduction

RNA is a molecule essential for life and has a variety of functions that makes it unique compared with DNA and protein. It is well established that one of its main roles is to carry the information encoded in the DNA (transcription) before protein synthesis (translation). However, several studies in the last decades have pointed to the existence of new RNA functions. The discovery of ribozymes in the 1980s [284-286] revealed an intrinsic catalytic activity of RNAs. Different RNAs are capable of catalysing different chemical reactions: small ribozymes can catalyse the formation and cleavage of a phosphodiester bond [287] and ribosomal RNA is involved in the formation of a peptide bond [288]. In addition, the discovery of riboswitches [289, 290] revealed a new mechanism that cells use to regulate gene expression. These RNA sequences are capable of assuming multiple structures with distinct functional roles.

In this chapter, a small model system is investigated, a 20nt RNA molecule, which can form two different hairpin structures that coexist in thermodynamic equilibrium. The EVB-SBP approach was used to study the conformational free energy landscape relying on available kinetic and thermodynamic data. Biased simulations using a non-geometric energy gap reaction coordinate enable the study of molecular pathways, which connects the two conformations, potentially revealing different structural mechanisms. In particular, the structural and energetic properties of the transition state ensemble have been studied in detail. I also carried out, in collaboration, the experimental investigation of the conformational properties of this RNA by using NMR spectroscopy to support the theoretical model.

5.2 Riboswitches

Living organisms have to control the expression of thousands of genes in response to metabolic demands and environmental changes. Transcriptional attenuation control is a classical mechanism used in the cell to regulate the level of gene expression, where proteins act as sensors of changes in the metabolite concentration [291, 292]. However, recent findings have shown that mRNAs can also respond to chemical and physical stimuli in order to control genes expression. These mRNA control systems, also known as "riboswitches", have been identified in prokaryotes and some eukaryotes [27, 28]. Riboswitches sense various metabolites, critical for fundamental biochemical processes, such as coenzyme B_{12} [290], thiamine pyrophosphate (TPP) [293], flavin mononucleotide (FMN) [289], S-adenosylmethionine (SAM) [294], lysine, guanine [295], and adenine [295]. It is interesting to note that genes controlled by ribowitches often encode proteins involved in the biosynthesis or transport of the metabolite being sensed [289, 293, 294]. This represents a simple and efficient mechanism of feedback inhibition whereby binding of the metabolite to the riboswitch decrease the expression of proteins directly involved in the metabolite synthesis. Riboswitches are generally located at the 5'-untranslated region (5'-UTR) of a particular mRNA and contain two main domains: an aptamer domain and an expression platform. The first is involved in the ligand binding, while the second undergoes an allosteric conformational change depending on the bound/unbound state of the aptamer domain [296].

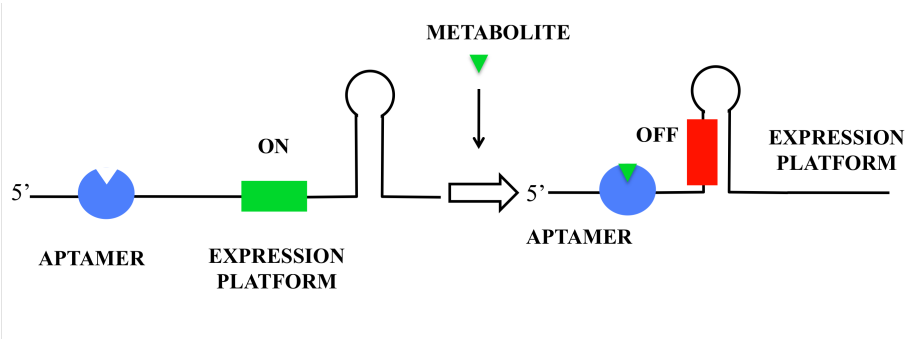

Figure 5.1 Schematic mechanism of a riboswitch function involved in gene control.

Each natural aptamer constitutes an extremely precise molecular sensor, able to recognize specific metabolites. One of the unique features of riboswitches is their enormous affinity and selectivity in binding their target molecules [297]. The structural

details of riboswitches have recently been revealed [298-300]. Experimental evidence also shows that the aptamer domain folds independently of the expression platform [301]. While the sequence of the aptamer domain has been conserved throughout evolution due to the fact that metabolites remain unchanged, heterogeneity is found in the expression platform domain sequence. The expression platform has the ability to change its secondary structure, an essential role for riboswitch functioning. Furthermore, riboswitches function by different mechanisms to modulate gene expression: control of the efficiency of translation initiation [290, 302]; control of the trancription elongation of mRNA [303]; control of the splicing of mRNA transcripts [303]. In order to fully understand riboswitch functioning, it is important to link the interaction between the ligand and the aptamer and the conformational change of the expression platform. Without aiming to fully understand the riboswitch function, here I focus on the RNA interconversion mechanism, by studying a bistable RNA model system.

5.3 Bistable RNA sequence

Some RNA can assume multiple stable structures with different functional roles. These sequences may be rationally designed to give two independent structures which either co-exist in thermodynamic equilibrium or can be triggered to change conformation in response to external perturbations, such as binding of a ligand or change in pH, temperature or salt concentration. Previous studies have demonstrated that even small RNAs containing only 20-30 nucleotides can be stable in different conformations at equilibrium [304]. I have investigated a 20nt bistable RNA sequence, which is known to adopt two coexisting hairpin loop structures, herein called A and B forms [305]. The specific 20nt RNA sequence studied here is 5'-r[GACCGGAAGGUCCGCCUUCC]-3' and the two hairpin structures (Fig. 5.2) differ in the number and identity of the constituting base pairs: A form has 4 base pairs, while the B form 6 base pairs [304]. This sequence has been appositely designed to give two competing structures. The 5'-end nucleotide sequence (GACC) of one hairpin and the 3'-end nucleotide sequence of the second hairpin (CGCCUUCC) compete for base pairing with the central sequence GGAAGGUC.

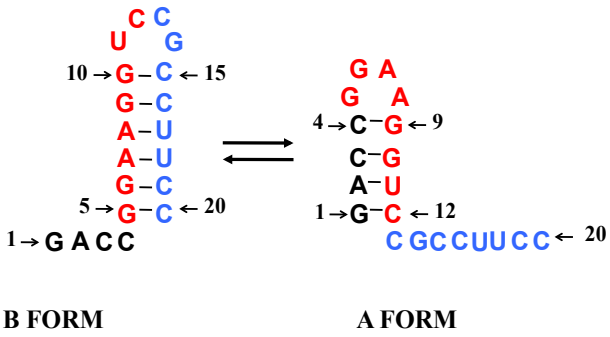

Figure 5.2 Two alternative, stable forms of a designed 20nt RNA sequence. The 5'-end (black) competes with the 3'-end (blue) for the base pairing with the central sequence (red).

The GGAA and UCCG tetraloop belong, respectively, to the GNRA and UNCG class (where N is any ribonucleotide and R is a ribonucleotide with a purine base), which together with the CUUG tetraloop account for almost 70% of known tetraloop structures. RNA tetraloops, despite their simplicity, play a key role in several processes ranging from RNA folding [306] and mediating tertiary interactions [307, 308] to providing recognition sites for proteins [309].

5.3.1 Previous experimental data

Experimental investigations using NMR spectroscopy have been conducted on the 20nt RNA sequence [304, 305]. First, conformational equilibrium between the A and B forms has been established [304]. The equilibrium was shifted toward the B form with a ratio between A and B of 25:75 at 298 K. Subsequently, kinetic parameters corresponding to the hairpin interconversion were measured by following the imino proton signal intensities of U11/G9 and U17/G10 [305]. In particular, the 20nt RNA sequence was modified to include a photolabile group on G6, which forms a base pair in form B but not in form A. As a result, the system was "caged" in the A form, preventing the A → B interconversion. A series of NMR spectra were then collected before and after photolysis until equilibrium was reached, as a function of temperature and concentration.

The equilibrium constant (K) was obtained from the relative intensities of the imino proton signals, providing the free energy change for the interconversion A → B: ΔG=-0.8 kcal mol^{-1} [305]. From the temperature dependence of K, using the standard van't Hoff equation, enthalpy and entropy values were obtained: ΔH=-5.9 ± 0.5 kcal mol^{-1} and ΔS=-17 ± 1.9 cal mol^{-1}K^{-1} [305]. Although the B conformation is more stable,

it is entropically less favourable due to the longer helix (6 base pairs) and the shorter single strand (4 residues). The rate constants were measured at different temperatures for both forward and backward transitions. These values range between $k_{(A\rightarrow B)}$=0.0136 s^{-1} at 283 K to $k_{(A\rightarrow B)}$=0.0132 s^{-1} at 298 K for the forward reaction and between $k_{(B\rightarrow A)}$=0.0019 s^{-1} at 283 K to $k_{(B\rightarrow A)}$=0.031 s^{-1} at 298 K for the backward reaction. Using the Arrhenius equation, activation enthalpy was obtained: $\Delta H^{\ddagger}_{(A\rightarrow B)}$=25.5 kcal mol^{-1} and $\Delta H^{\ddagger}_{(B\rightarrow A)}$=30.6 kcal mol^{-1} [305]. The authors proposed that since the activation enthalpy amounts to about half of the base pairing enthalpy of 55 kcal mol^{-1}, determined by thermal denaturation studies of truncated RNA hairpins [304], the transition structure retains about half of the initial number of base pairs.

In addition, imino proton exchange rates were extracted from NOESY cross peaks [305]. These values were compared with those of the truncated RNA hairpins that are conformationally locked to represent either fold A or fold B. The authors claim that a general increase in water exchange rates of the 20nt RNA compared to those of the truncated hairpin occur, in particular for those base pairs that are closer to the loop region.

Based on these findings, a switching mechanism was proposed where the disruption of base pairs close to the loop (unfolding) and formation of new base pairs between the loop region and the single strand (refolding) take place simultaneously in an associative fashion [305]. This involves a transition structure with 6 base pairs: 4 base pairs of the B form and 2 base pairs of the A form. It is however important to point out that this structural hypothesis of the mechanism is based on kinetic and thermodynamic data only with no direct evidence.

5.4 Two-dimensional modelling of bistable RNA

RNA folds according to a hierarchical model, where the formation of secondary structural elements, which represent the bottleneck of the process, precedes tertiary assembly [310]. Therefore the first step toward the understanding of the three-dimensional structure and function of RNA is the study of stable secondary structure elements. The analysis of secondary structure motifs formed in the bistable RNA sequence and possible interconversion mechanisms are described in the following.

5.4.1 Secondary structure prediction

Several theoretical and computational models have been proposed to predict the formation of secondary structure elements in RNA sequences. It is assumed that the total free energy of the structure can be calculated as the sum of the free energies of the individual base pair stacks and loops. The empirical thermodynamic parameters used are known as "Turner rules" [311-313]. Most approaches are based on energy minimisation using the "nearest neighbour" model [314-316] where the stability of a helical segment is calculated as a pairwise sum of the free energy contributions from all base pair stacks, eg. GA/CU. In addition, the hairpin loop energy is calculated as the sum of an adverse entropic contribution depending on the loop size, and a favourable stacking term between the closing base pair of the hairpin loop and the adjacent pair. These numerical approaches have been interfaced by a number of web servers, allowing an easy investigation of RNA secondary structural elements, such as mfold [317], Vienna RNA [318], Sfold [319], and RNA shapes [320].

Here I use the *mfold* web server to calculate the stable secondary structures of the model RNA sequence at 310 K [317]. The RNA sequence was calculated to fold into two free energy minima corresponding to the A and B forms (Table 5.1).

Table 5.1 The free energy contributions to secondary structures in the B and A forms calculated by *mfold* [317]

B-form		
Structural element	**δG (kcal mol $^{-1}$)**	**Information**
Single strand	-0.30	4 ss bases and 1 closing base pair
Stack	-3.30	External closing pair G^5-C^{20}
Stack	-2.40	External closing pair G^6-C^{19}
Stack	-0.90	External closing pair A^7-U^{18}
Stack	-2.10	External closing pair A^8-U^{17}
Stack	-3.30	External closing pair G^9-C^{16}
Hairpin loop	3.40	Closing pair G^{10}-C^{15}
Total free energy	**-8.9**	

A-form		
Structural element	**δG (kcal mol $^{-1}$)**	**Information**
Single strand	-0.80	8 ss bases & 1 closing base pair
Stack	-2.40	External closing pair G^1-C^{12}
Stack	-2.20	External closing pair A^2-U^{11}
Stack	-3.30	External closing pair C^3-G^{10}
Hairpin loop	0.40	Closing pair C^4-G^9
Total free energy	**-8.3**	

In addition, assuming a two-state model (folded/unfolded), an estimate of the melting temperature (T_m) can be calculated using the enthalpy and entropy contribution to free energy. Since at the melting temperature $\Delta G°=0$, T_m can be calculated as $T_m=\Delta H°/\Delta S°$ (Table 5.2).

Table 5.2 Thermodynamic data for B and A forms calculated using the *mfold* server [317]

	$\Delta G°$ (kcal mol^{-1})	$\Delta H°$ (kcal mol^{-1})	$\Delta S°$ (cal mol^{-1}K^{-1})	T_m(K)
B Form	-8.9	-60.9	-167.6	363.0
A form	-8.3	-54.2	-147.9	359.3

These results provide an estimate of the relative stabilities of the two conformations at the secondary structure level. Consequently, this 20nt RNA sequence can form both the A and B conformations, the B form being more stable than the A form, consistent with experimental information.

5.4.2 Interconversion mechanism at the secondary structure level

The secondary structure prediction algorithms discussed above provide a useful tool to obtain free energy minima given a RNA sequence. Following these models, several approaches have been introduced to also predict the folding path of nucleic acids and provide insight into the kinetics of the process at the secondary structure level.

I employed the program Kinefold [321] to study the possible interconversion mechanisms between the stable forms of the bistable RNA. Kinefold performs stochastic folding simulations of nucleic acids using the *dynamic folding algorithm* [322]. The folding pathway is represented by successive steps of nucleation and dissociation of helix regions using a kinetic Monte Carlo procedure. For each new step, the free energy of possible new base pairs are evaluated and compared to the previous state. The transition rates are then recalculated and new Monte Carlo movement is performed. This process is repeated until the time-average distribution of configurations appears stationary. To avoid kinetic traps and increase efficiently of the code, a clustering algorithm is used in Kinefold [321].

The analysis of the 20nt RNA sequence is performed using the renaturation procedure, where the starting structure is a single strand which is instantaneously cooled to 300 K. I have performed 100 refolding simulations with different random seeds from a completely denatured single strand RNA that has a relative free energy of

0 kcal mol^{-1}. In 37 trajectories, the system relaxes within 0.003 ms to the stable hairpin form B with 6 base pairs at -8.9 kcal mol^{-1} (Fig. 5.3, Path-1). In all other cases, there are intermediate states along the folding trajectory to form B. In 36 cases the folding takes place over 0.23-0.55 ms; the first intermediate is a hairpin motif with three base pairs at -3.4 kcal mol^{-1} (0.003 ms) and the second intermediate predominantly at -4.9 kcal mol^{-1} (0.054 ms) is a pseudoknot structure with two intercalated helices with 3 base pairs each (Fig. 5.3, path-2).

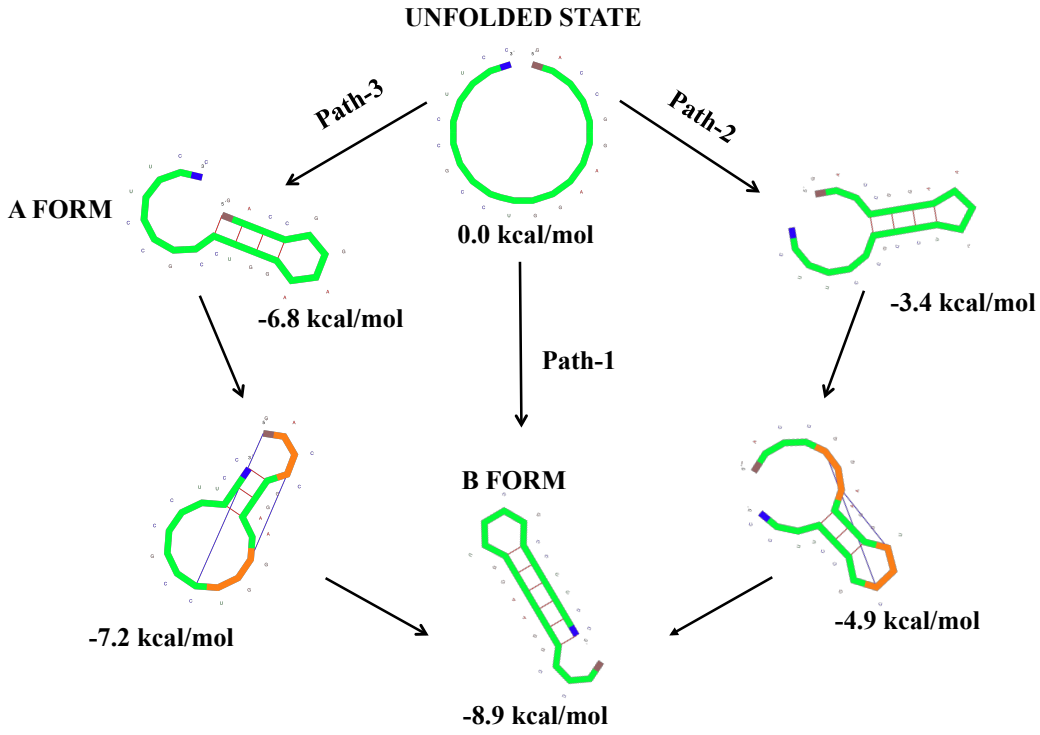

Figure 5.3 Refolding pathways calculated using Kinefold [321]. 5'-end and 3'-end are depicted, in gray and blue, respectively. Orange segments represent intercalated helices, delimited by blue lines, base pairs are depicted by red lines.

While the first intermediate structure includes base pairs that are not native to either A or B hairpin forms, in the second pseudoknot structure, 3 base pairs are native and 3 base pairs are not native. Pseudoknot structures are characterised by two hairpin loops in which the stems are intercalated [323, 324]. These structures however do not represent true topological knots of the biopolymer and hence it can unfold. The prediction of such entangled structures is not trivial and only a handful of algorithms are available today for this purpose. All other trajectories require orders of magnitude longer times (12-13 ms) to fold to the stable hairpin form B. These pathways (with one exception) fold first to hairpin form A at -6.8 kcal mol^{-1} within 0.003 ms, then to

another pseudoknot structure with a surprisingly low free energy of -7.2 kcal mol^{-1} at 0.038 ms (Fig. 5.3, path-3). In this pseudoknot base pairs characteristic of both A and B forms are present simultaneously: 4 base pairs of the A form and 3 base pairs of the B form coexist. The enhanced stability of this species is also shown by the fact that only after several hundred moves and over 10 ms simulation time the system can reach hairpin form B.

There are interesting conclusions to be drawn from the secondary structure folding pathways. First, a pseudoknot with native base pairs appears to lie on the folding pathway between hairpin folds A and B. The pseudoknot also represents an intermediate structure with an intermediate energy between folds A and B. Second, both hairpin folds A and B can be directly accessed from the denatured single-stranded RNA. Thus there are two extreme cases for the transition pathway between form B and form A of the RNA hairpin: i) an associative mechanism characterized by a pseudoknot-type transition state, and ii) a dissociative mechanism where the transition state is a completely unfolded single strand. The secondary structure predictions described above can provide useful indications on the possible mechanism of the interconversion pathway, however, molecular simulation approaches are needed to describe the switching mechanism in detail. In the following sections, experimental and computational investigations are described to shed light on the interconversion mechanism of the bistable RNA sequence.

5.5 Three-dimensional modelling of stable RNA structures

As described in Chapter 3, the prerequisite of the application of the EVB-SBP method is the knowledge of the structural endpoints involved in the conformational transition. Due to the absence of the experimental structure of either conformer in the literature, I constructed these structures by assembling a canonical A-form duplex, an experimental tetraloop structure and a single strand (Fig. 5.4).

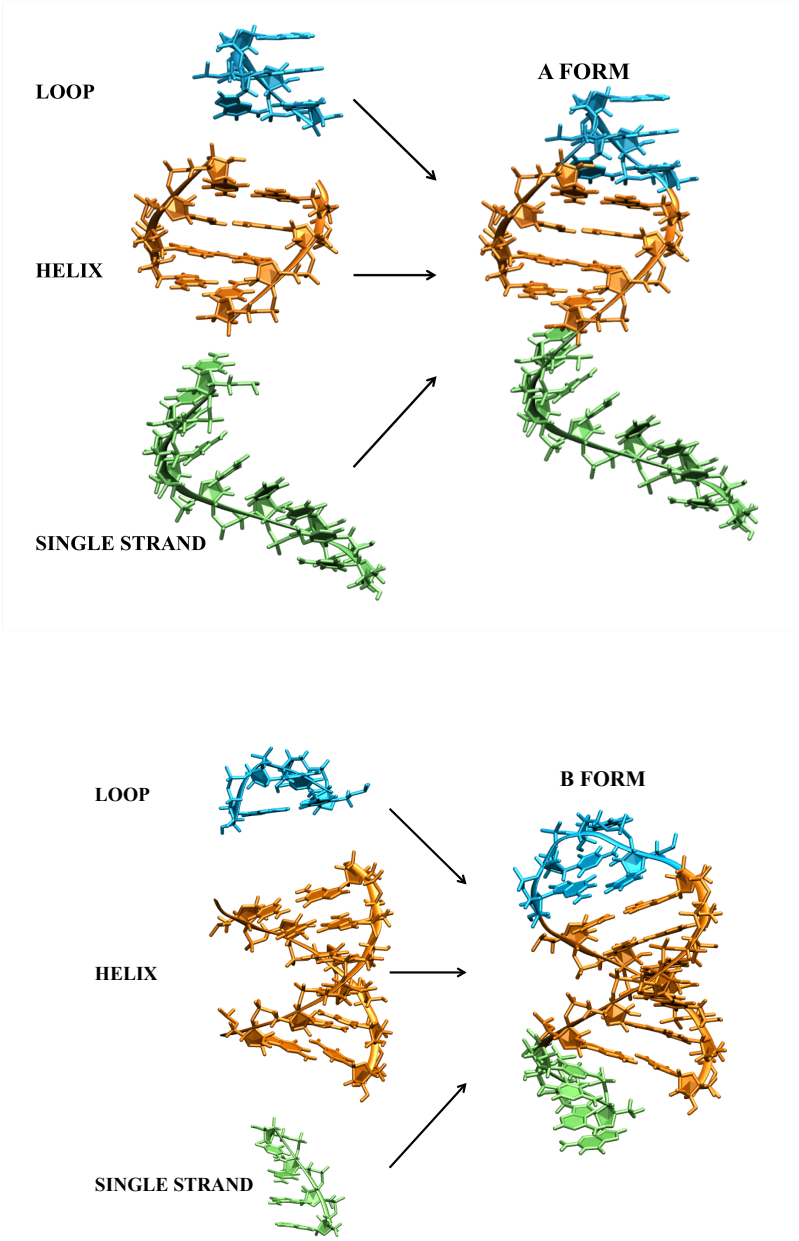

Figure 5.4 Schematic description of model building for the A form (top) and B form (bottom) RNA hairpin loops.

For the construction of the helical stem and the single stranded portion of the structure, the *nucgen* program (part of the Amber package [243]) was used with standard fibre-diffraction parameters. Structures of the A and B forms involve different tetraloop sequences: GGAA and UCCG, respectively. The GGAA tetraloop (for the A

form) has been modeled using portion of the X-ray structure of the 7S.S SRP RNA complex [325] (PDB 1LNG), while the UCCG tetraloop (for the B form) has been modeled using the the X-ray structure of a UUCG tetraloop of the 16S ribosomal fragment [326] (PDB 1F7Y) and manually introducing a U → C mutation. The energy-minimised structures represent the conformational endpoints of the interconversion process and, in particular, serve as reference structures for the definition of the structure-based potential.

In order to gain insight into the structural and energetic stability of the model, both A and B structures, have first been studied by molecular dynamics simulations. The RNA was solvated with ~8125 TIP3P [163] water molecules and 19 K^+ counterions in a truncated octahedral box (~282,203 $Å^3$) with periodic boundaries, which allowed for ~12 Å shell of water molecules between the solute and the edge of the periodic box. Molecular dynamics simulations were carried out at constant temperature (300 K) and pressure (1 bar) using the parm99 of the Amber force field [153]. An integration time step of 2 fs was used and all bond lengths involving hydrogens were constrained using SHAKE [327]. Long-range interactions were treated using the PME approach [160] with an 8 Å direct space cut-off. Both starting RNA structures were minimized using steepest descent conjugate gradient methods (see Chapter 2). The system was gradually heated at constant volume (50 ps) using positional restraint (k=25 kcal mol^{-1} $Å^{-2}$) and then switched to constant pressure (500 ps) and decreased the positional restraint on the solute atoms. After the equilibration procedure, simulations were continued for 70 ns. The purpose of these simulations was to determine the structural stability of the model structures built and allow these to relax to a free-energy minimum.

A structural analysis of the helix, loop and single strand has been carried out along the trajectories by calculating the all atom RMSD values for the sub-structures separately (Fig. 5.5).

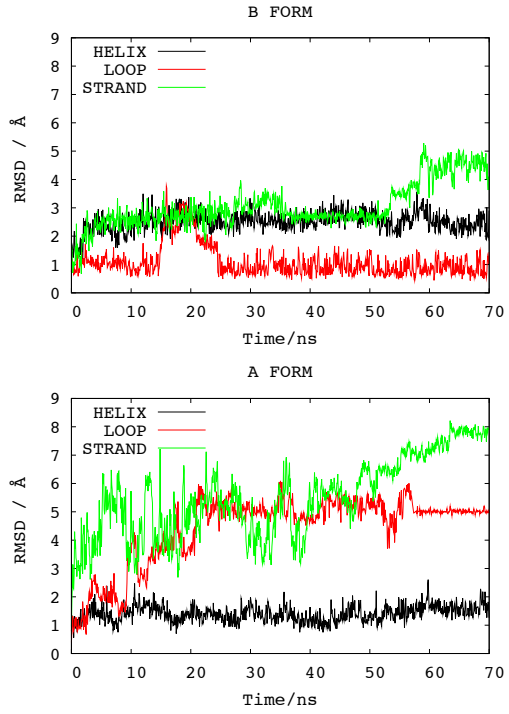

Figure 5.5 Root mean square deviation calculated for all atoms forming the helix, the loop and the single strand is shown along the trajectory for the B form (top) and A form (bottom).

The tetraloop in the B form (5'-UCCG-3'), shows an average RMSD of 1.0 ± 0.2 Å, and undergoes a local conformational change (RMSD ~2.2 ± 0.6 Å) in the time interval between 15 and 25 ns. A visual analysis reveals the flip of base G14 out of the loop (Fig. 5.6). In addition, in line with a recent study [161], the H-bond U11(O2')-G14(O6), a key interaction for the stability of the UUCG loop [328], is lost during the first nanosecond and replaced by the U11(O2')-C12(O5') H-bond. Such structural changes have been described [161] as a result of force field inaccuracies. The A form tetraloop (5'-GGAA-3') is relatively stable in the first 10 ns with an average RMSD of 1.6 ± 0.6 Å. However, after 10 ns, the initial conformation, stabilized by stacking interactions between the three bases G6/A7/A8, is disrupted with the three bases fully exposed to the solvent environment (Fig. 5.6). This results in an increase in RMSD to ~5 Å that is stabilised by stacking interactions of G5 and G6 bases.

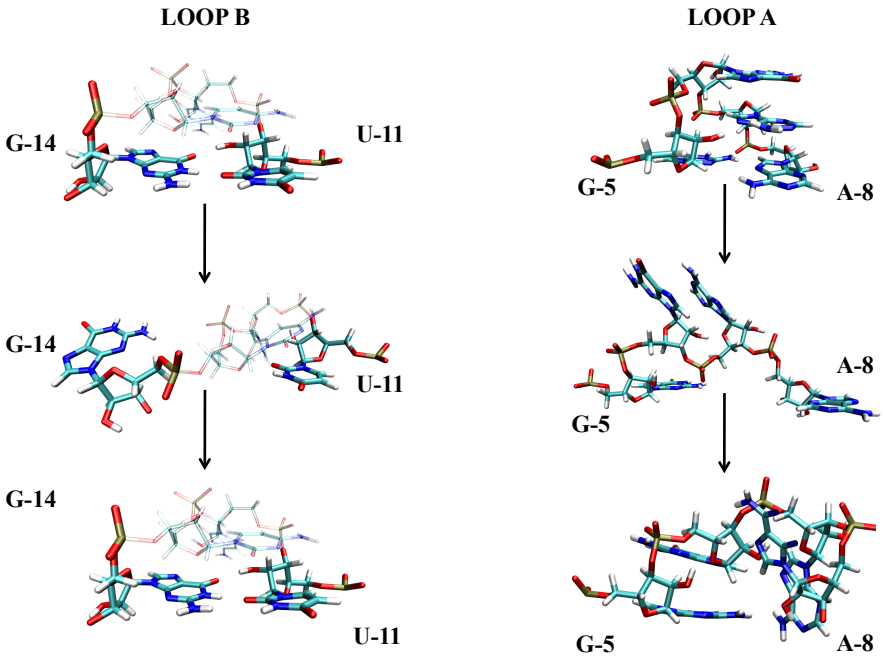

Figure 5.6 Representative tetraloop structures along the molecular dynamics trajectories. The flip of base G14 is shown in loop B (left); overall structural changes in loop A (right).

The all atom RMSD calculated for the helix reaches an average value of 1.4 ± 0.3 Å for the A form and 2.5 ± 0.4 Å for the B form. The slightly higher RMSD in the B form is mainly due to the opening of the terminal base pair G5-C20. In addition, the flipping of the base G14 out of the loop induces a slight "stretch" of the helix, as evidenced by a slight increase in the end to end helix distance from 14.0 ± 1.0 Å to 16.8 ± 1.0 Å and a decrease of the average helical twist from 28.5° ± 1.5 to 24.7° ± 1.5, with 31.1° being the reference value from experimental structures [269]. Similar underwound "ladder-like" RNA helix deformations were also reported in a recent study on tens-of-nanosecond time scale and have been attributed to force field artifacts [329].

The single strand in the B form, built in a helical conformation, rearranges toward a more compact state during the simulation. The center of mass distance was calculated between terminal residues G1 and C20 along the trajectory (Fig. 5.7). Initially, when the single strand is in an extended helical conformation, the distance fluctuates around ~12.5 Å. However, when the strand folds to a compact state, it decreases to 8 Å. In this conformation the single strand is in the proximity of the helix, mainly stabilized by interaction between C3(N3) with C20(O3') and C4(N3) with C20(O2'). A similar scenario is observed for the A form. The G1-C20 distance in the longer single strand

(5'-CGCCUUCC-3') is reduced from 36 Å to ~5 Å in the final structure (Fig. 5.7). A visual analysis reveals that in this conformation residue C20 and C13 form a non-canonical base pair, also known as C-C N3-amino symmetric [19], which is stabilized by stacking interactions with the first base pair G1-C12.

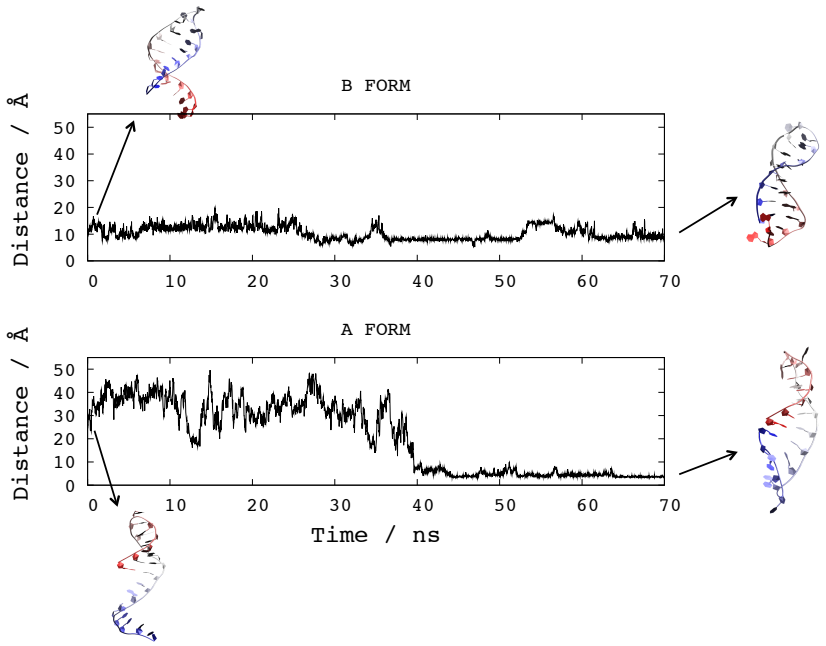

Figure 5.7 Time series of distance calculated between the centres of mass of bases G1 and C20 for the B form (top) and A form (bottom).

In summary, in line with previous studies, the present analysis points to some force field inaccuracies, which result in structural rearrangements in the simulations. It should be noted that an improvement to the current Amber force field has recently been proposed [161]. The use of such a force field would probably improve the quality of three-dimensional model building of RNA, however, it was not yet available when these simulations were performed. In the following, simulation data from the first 5ns are used in the comparison with the structure-based potential where no major structural changes were observed.

5.6 EVB-SBP simulation of bistable RNA

I will describe here the main steps toward the application of the EVB-SBP method to study the bistable RNA sequence. First, the parameterisation of the structure-based

potential (SBP) for the individual conformations is described, followed by the parameterisation of the combined free energy surface, and finally biased simulations on the parameterised free energy surface are shown.

5.6.1 Structure-based potential

Based on previous results [238], a cut-off for the non-bonded interactions of 4.5 Å and an energy rescaling factor S=2.5 are used. In order to successfully apply the method to the RNA model, a modification of the method has been introduced. As described in Chapter 3, the native interactions in the SBP are described by classical Lennard-Jones 12-6 potentials. While the equilibrium distance for each contact was obtained from the van der Waals radii of the Amber parm99 force field in Chapter 4, the equilibrium value for each native inter-atomic distance will be taken directly from the reference structure in this application. This modification should ensure that the native hairpin structures are better reproduced. Nevertheless, due to the free rotation around the glycosidic bond of unpaired bases, also described in Chapter 4, inter-conversion trajectories result in both *anti* and *syn* bases in the final structures. Although base pairs with *syn* bases in the helix result in higher free energy structures, there is a significant probability of incorrect closing of at least one of the 10 different base pairs in the two RNA hairpins. To control the correct closing to *anti* conformations, a native intra-residue interaction was introduced between N3 (purines) or O2 (pyrimidines) and the sugar O4' atoms. These distances are around ~4 Å in the *anti* conformation and decrease to ~3 Å in the *syn* conformation. Consequently, introducing a LJ potential with a minimum at 4 Å will prevent the transition to *syn* due to the repulsive potential. It is remarked again that the χ angle has recently been reparameterised in the Amber parm99 force field [161] and may eliminate the need for the correction introduced here.

Although a similar number of native contacts (N_A=1065; N_B=1055) exists in the A and B forms, a notable difference is seen in the number of contacts present in the B-loop (UCCG) and the A-loop (GGAA): while only 96 native contacts are present in the B-loop, a total of 197 contacts are present in the A-loop. This is mainly due to the different structural arrangements of the two central bases. In the B form, bases C12 and C13 lie on opposite sides of the loop resulting in a few number of contacts; in the A form bases G6/A7/A8 are stacked, resulting in a higher number of native contacts. As a

consequence, A-loop is considerably more stable in the standard parameterisation, which introduces hysteresis in the calculated free energy profile sampled on the ns-μs timescales. As it will be described below, tetraloops play a crucial role in initiating the zipping mechanism of helix formation. Since it was considered that the formation of the B-loop is significantly hindered by the relatively low number of native contacts, the rescaling factor of S=2.5 has been increased to S=6.25 for native interactions in the B-loop to provide reversible free energy profiles.

Finally, it has to be noted that the use of exact interatomic distances in this model requires a particular treatment of those native contacts which are present in both conformations (see Chapter 3). Significant energy fluctuations were observed (Fig. 5.8) in Langevin dynamics simulations (25 ns) performed at 300 K and low friction (γ=5 ps^{-1}). Since the value of the diabatic energy is used in the energy gap reaction coordinate, such fluctuations are not satisfactory. These energy fluctuations are mainly due to the fact that LJ potential has a very stiff repulsive part at short distances and small changes in structure results in a large change in the corresponding energy. In particular, the increased effect for state A in the first basin is probably due to atomic contact distances which undergo large fluctuations. The introduction of a flat bottom LJ potential for such native contacts present, alleviates this problem (Fig. 5.8).

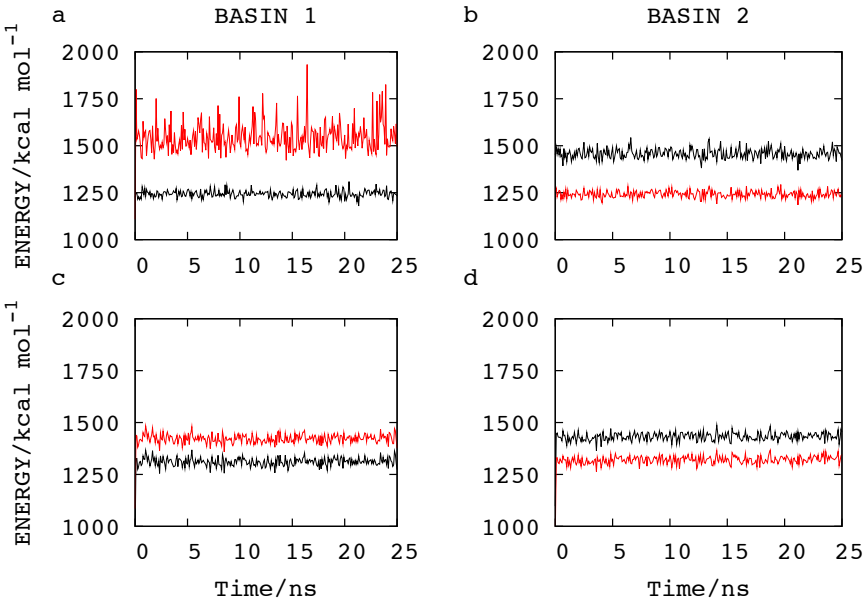

Figure 5.8 Time series of the diabatic energy of RNA hairpin forms B (black) and A (red). Basin 1 corresponds to a geometry where form B is stable; basin 2 corresponds to a geometry where form A is stable. Top panels (a) and (b) refer to the 'original' model, while bottom panels (c) and (d) show the results of simulations where the flat bottom potential for native contacts in both states was introduced.

Results from simulations using the parameterised SBP described above were compared with those from 5 ns all-atom force field simulations as reference data (see section 5.5). Structural stability and fluctuations were compared using RMSD metrics. All-atom RMSD values for the helix residues in both models show that the difference between SBP and FF is practically indistinguishable (Fig. 5.9). In the case of the A form RNA structure, the average RMSD using FF and SBP is 1.3 Å ± 0.3 and 1.0 Å ± 0.2, respectively; in the case of B form RNA structure, those are 2.0 Å ± 0.3 and 1.8 Å ± 0.5 for FF and SBP, respectively.

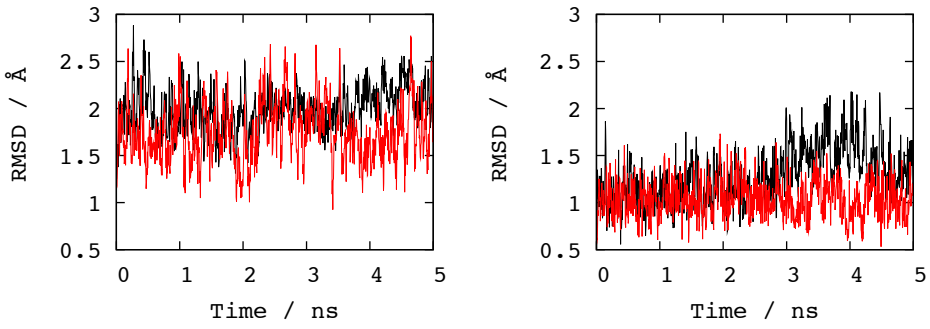

Figure 5.9 Root mean square deviation of the atomic positions for helix B (left panel) and helix A (right panel) using 5 ns MD simulations with the parm99 Amber force field (black) and SBP (red).

The average per residue fluctuation calculated from simulated trajectories with SBP is, however, lower ($RMSF_A$=0.5 Å ± 0.2; $RMSF_B$=0.7 Å ± 0.2) compared with that with force field ($RMSF_A$=1.3 Å ± 0.2; $RMSF_B$=1.5 Å ± 0.3).

5.6.2 Parameterisation of the EVB-SBP free energy surface

The prerequisite to study the interconversion of the bistable RNA by the EVB-SBP method is that the parameterised free energy surface matches critical experimental data. For this purpose, the EVB parameterisation introduces optimal diabatic state shift ($\Delta\alpha$) and coupling constant (V_{12}) terms. Experimental data [305] provide directly the free energy difference of ΔG (A→B) ≈ -0.8 kcal mol^{-1} between A and B forms, and the activation free energy can be calculated from experimental rate constants measured at different temperatures. Assuming transition state theory [330] holds, rate constant is

directly related to the activation free energy of the process by the Eyring-Polanyi equation:

$$k = \kappa \frac{k_B}{h} * e^{-\frac{\Delta G^{\ddagger}}{RT}} \tag{5.1}$$

where k is the reaction rate constant, κ the transmission coefficient (here taken to be unity), k_B is the Boltzmann constant, h the Planck constant, $\Delta G^{\ddagger}$ the Gibbs free energy of activation, R the gas constant and T the absolute temperature. Moreover, using the linear form of the Eyring-Polanyi equation, one can calculate both enthalpy and entropy of activation:

$$\ln \frac{k}{T} = \frac{-\Delta H^{\ddagger}}{R} \frac{1}{T} + \ln \frac{k_B}{h} + \frac{-\Delta S^{\ddagger}}{R} \tag{5.2}$$

where, $\Delta H^{\ddagger}$ is the entalphy of activation and $\Delta S^{\ddagger}$ the entropy of activation. The thermodynamic parameters calculated from best-fit line obtained using four experimental data points [305] are reported in Table 5.3.

Table 5.3 Activation parameters for the interconversion of the bistable RNA calculated from experimental data. [1]

	$\Delta H^{\ddagger}$	$\Delta S^{\ddagger}$	$\Delta G^{\ddagger}$(298K)	$\Delta G^{\ddagger}$(283K)
Transition A→B	24.96	20.93	18.64	18.94
Transition B → A	30.78	37.83	19.49	20.04

[1] Enthalpy and free energy are reported in kcal mol^{-1}, and entropy in cal mol^{-1} K^{-1}.

It is notable that the experimental activation free energy values are significantly smaller compared to the activation enthalpy, due to an increase in entropy in the transition state with respect to the stable states. Indeed, the activation entropy change is positive in both directions, but higher in the B→A transition, due to the longer helical stem and shorter single strand segments in the B form.

In order to match these experimental thermodynamic and activation data, the EVB parameterisation procedure, consisting of tuning the diabatic state shift and the coupling element, was employed. Simulations were performed according to the method described

in Section 5.6.3. Initially, the diabatic state shift and the coupling constant were set respectively to $\Delta\alpha_{12}$=0 kcal mol^{-1} and V_{12}=10 kcal mol^{-1}. The resulting free energy difference $\Delta G_{B\rightarrow A}$~10 kcal mol^{-1} and the activation free energy $\Delta G^{\ddagger}_{B\rightarrow A}$~50 kcal mol^{-1} were both higher compared to the reference values. Based on these preliminary results, the coupling constant was increased to V_{12}=45 kcal mol^{-1}, in order to lower the activation free energy ($\Delta G^{\ddagger}_{B\rightarrow A}$~26 kcal mol^{-1}), and the diabatic state shift was modified to $\Delta\alpha_{12}$=-10 kcal mol^{-1} with the aim of reducing the free energy difference ($\Delta G_{B\rightarrow A}$~3 kcal mol^{-1}). With the same intent, a further modification of the coupling constant to V_{12}=80 kcal mol^{-1} and of the diabatic state shift to $\Delta\alpha_{12}$=-60 kcal mol^{-1} was applied. At this point, though the correct activation free energy ($\Delta G^{\ddagger}_{B\rightarrow A}$~19 kcal mol^{-1}) was achieved, the resulting free energy difference ($\Delta G_{B\rightarrow A}$~0 kcal mol^{-1}) was still not matching the reference value. In the final step the energy of form A was increased by reducing the diabatic state shift to $\Delta\alpha_{12}$=-58 kcal mol $^{-1}$. In conclusion experimental data can be best matched in the EVB-SBP simulations using diabatic state shift, $\Delta\alpha_{12}$=-58 kcal mol $^{-1}$ and coupling constant, V_{12}=80 kcal mol^{-1}.

An alternative way of parameterising the free energy surface can also be considered for future applications. This approach consists of performing an initial biased simulation in order to calculate the diabatic state energies, the ground state energy and the free energy along the transition. An estimate of the entropic term can also be computed from the difference between the ground state energy and the free energy. The diabatic state energies can then be used to recalculate analytically the EVB ground state energy, using different combinations of V_{12} and $\Delta\alpha_{12}$. Assuming that the entropy is constant for different combinations of EVB parameters, one can obtain the new free energy by combining the new ground state energy calculated analytically and the entropy estimate. The main advantage of this approach is to avoid the trial and error procedure, described above, where multiple and time consuming biased simulations have to be performed.

5.6.3 Simulation on the parameterised free energy surface

The umbrella sampling technique in combination with the general energy gap reaction coordinate ($\Delta\varepsilon$) is used to drive the system between the A form and B form RNA structures. An ensemble of 50 trajectories was generated for both transition directions: 25 trajectories for B→A and 25 trajectories for A→B. Umbrella windows are

positioned along the reaction coordinate with a step of $\Delta\varepsilon$=1 kcal mol^{-1} sampled for 1 ns. Individual simulations are 326 ns long, with a total combined simulation time of 16.3 μs. The large body of simulation data allows for detailed analysis of the interconversion mechanism between the two conformations. The results from each biased sampling window were then combined and unbiased by the weighted histogram analysis method (WHAM) [211](see Chapter 2).

As previously noted [238], due to the functional form of the LJ potential, the transition state does not correspond to $\Delta\varepsilon$=0 kcal mol^{-1} as is the case using harmonic potentials to describe native interactions. In the case of the bistable RNA, the energy gap ranges between $\Delta\varepsilon_1$ = −138 kcal mol^{-1}, corresponding to the B form minimum, and $\Delta\varepsilon_2$ = 188 kcal mol^{-1}, corresponding to the A form minimum (Fig. 5.10).

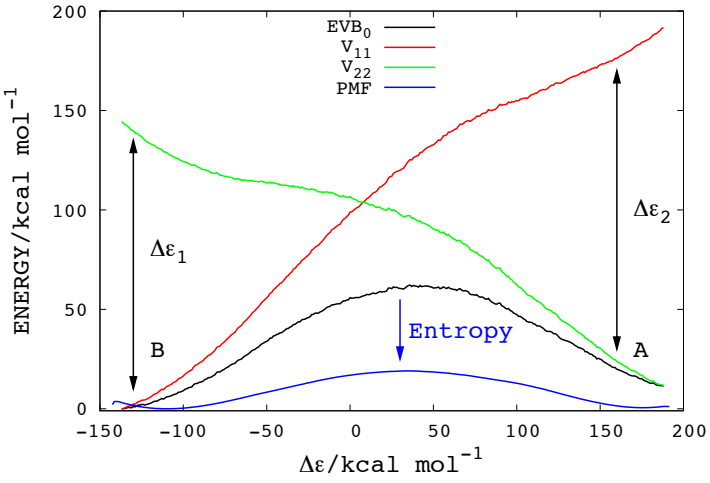

Figure 5.10 Energy profile of the RNA conformational change along the energy gap reaction coordinate. Diabatic state energies are shown in green and red, EVB ground state energy in black, and free energy profile in blue. Minima corresponding to forms A and B of the RNA are indicated. Results are averaged over 50 trajectories.

The comparison between the EVB ground state energy and the free energy profile allows for a quick estimate of the entropy contribution along the transition pathway. The free energy and the EVB ground state energy have been aligned with respect to the B form. The reason for this choice is that the entropy effect is expected to be higher in the A form due to the presence of a longer single strand (8 residues) compared to the B form (4 residues). It is notable that the largest entropy contribution is observed at the transition state, indicating a more heterogeneous structural ensemble than those of the stable states. Since only the experimental free energy data were used to parameterise the

EVB-SBP potential, the calculated activation enthalpy and entropy values along the transition pathway should be considered qualitative.

Individual trajectories show free energy variations of ± 3 kcal mol^{-1} around the average free energy, that reproduces the reference experimental data well. The calculated average activation free energy is $\Delta G^{\ddagger}_{B\rightarrow A}$=19.05 kcal mol^{-1} and the difference in free energy is $\Delta G_{B\rightarrow A}$=0.6 kcal mol^{-1} (Fig. 5.11).

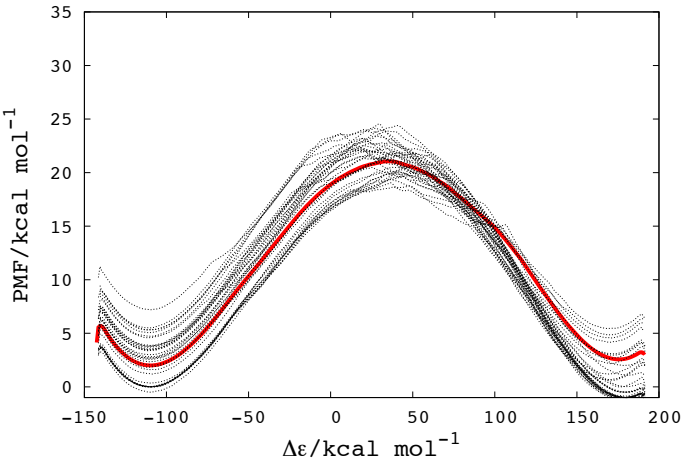

Figure 5.11 Free energy changes along the transition pathway between the stable RNA structural forms B and A calculated using the EVB-SBP method. Dotted lines show free energy profiles calculated from individual molecular trajectories; solid line represents the ensemble-averaged free energy profile.

Using the energy gap reaction coordinate no direct geometric restraint is imposed on the system, allowing an extended exploration of the conformational space. The energy gap serves as a good progress variable in that it clearly distinguishes between the initial and the final states without influencing the structural properties of the intermediate states. A detailed structural analysis, however, reveals different transition pathways between the A and B forms.

In order to depict the overall structural changes along the transition pathway, the radius of gyration (R_g) for the 20nt RNA is calculated from all simulations (Fig. 5.12). Data from multiple windows are merged together, and the probability distribution is calculated: Thirty consecutive windows, which correspond to an energy gap interval of 30 kcal mol^{-1}, provide a “macro-window”.

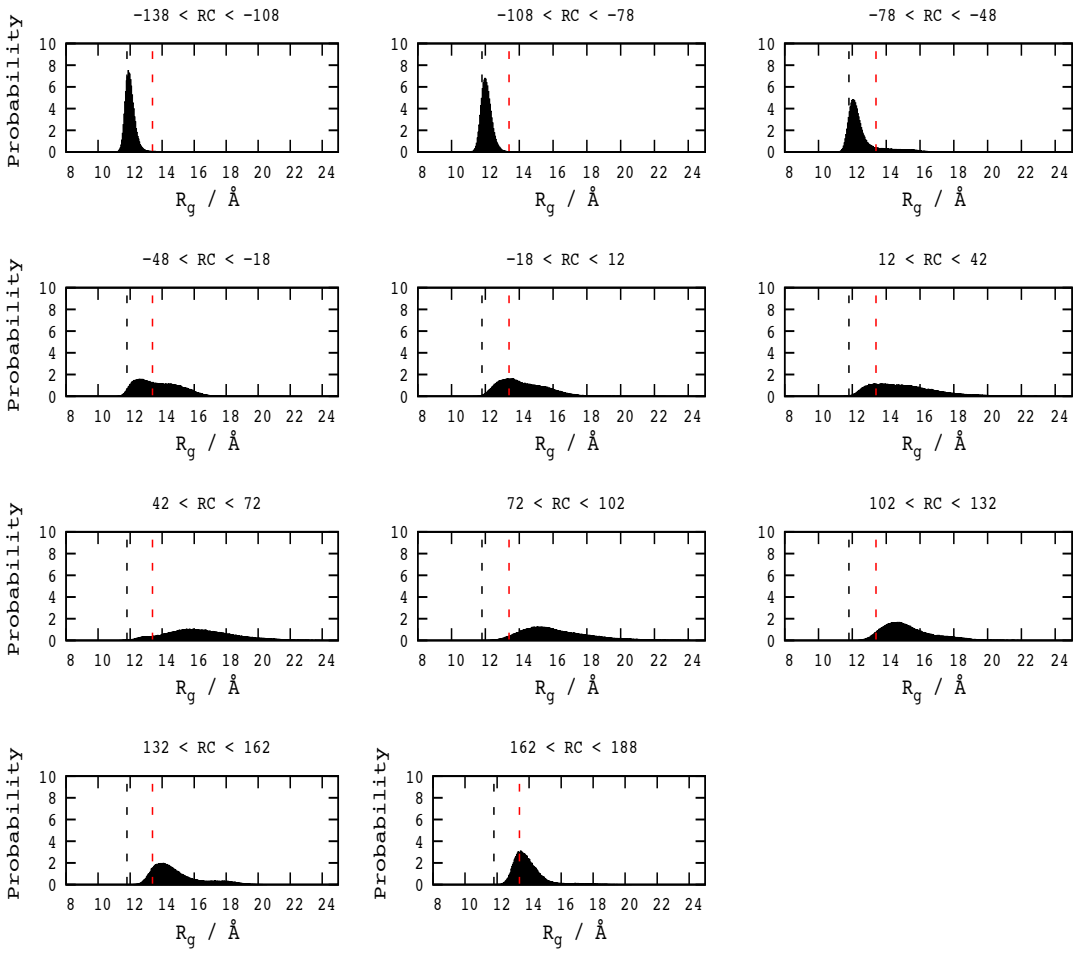

Figure 5.12 Probability distribution of the radius of gyration calculated in "macro-windows" generated from independent umbrella sampling simulations. Black vertical dashed line indicates the reference value for state B and red vertical dashed line that for state A. RC indicates the energy gap (Δε) reaction coordinate.

Reference values for the analysis are calculated from native state dynamics of the respective conformations. In the case of form B (-138 kcal mol^{-1} < Δε <-108 kcal mol^{-1}) the distribution of R_g is very tight, while in form A (162 kcal mol^{-1} < Δε < 188 kcal mol^{-1}) the distribution is considerably wider. This can be explained by the larger fluctuations of the longer single strand in form A compared with form B. The intermediate region (12 kcal mol^{-1} < Δε < 72 kcal mol^{-1}) shows a broad distribution of R_g, ranging from 12 to 24 Å, indicating a heterogeneous structural ensemble. The intermediate region, which is a potential transition state, exhibits longer average R_g with respect to either stable states and thus involves non-compact structures.

5.7 Free energy map and interconversion pathways

Driving the system using the energy gap reaction coordinate is an effective way to achieve an extensive conformational sampling but it does not relate energy changes to any structural information. In order to study the structural changes along the pathways and characterise the transition state ensemble, the free energy is mapped onto structural properties such as RMSD calculated with respect to both helices which, in this specific case, clearly distinguishes between initial and final states. I have calculated the potential of mean force as a function of the all-atom RMSD from the A form helix and the all-atom RMSD from the B form helix using the conformational ensemble sampled along the energy gap reaction coordinate. In order to extract this information a three-dimensional WHAM procedure was used [215, 281] where conformational statistics was binned along restrained Δε, and unrestrained RMSD (from A) and RMSD (from B) coordinates. Subsequently, the restrained coordinate (Δε) was integrated out to provide the two-dimensional free energy surface along the RMSD coordinates only. In the following, three examples are presented from individual trajectories to show structural variations in the transition pathways which describe the inter-conversion mechanism.

In the first case, the transition from the RNA hairpin form B to A proceeds via a highly-disordered, non-compact state (Fig. 5.13).

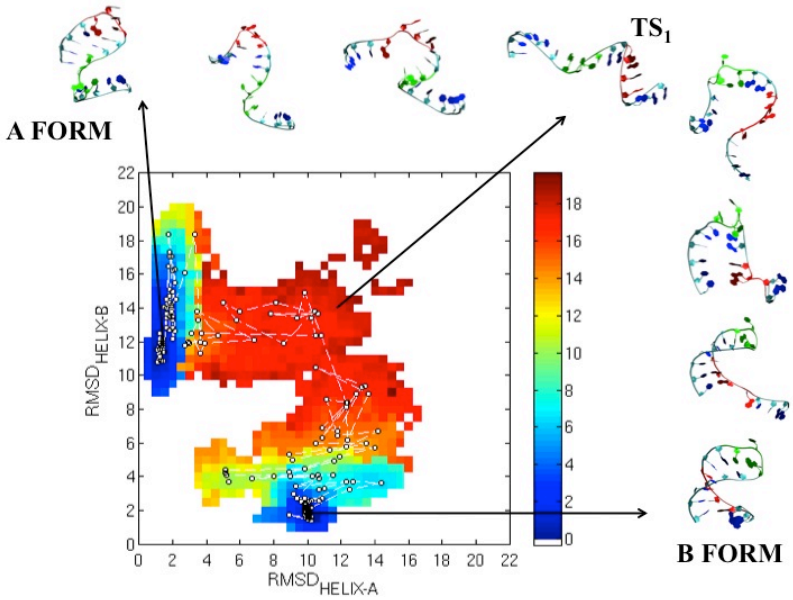

Figure 5.13 The free energy surface sampled by an individual trajectory started in form B (pathway 1) is shown as a function of the RMSD from the A form helix and from the B form helix. The white line shows the approximate path traversed during the simulation and the white points are obtained by averaging RMSD values in each window. Representative structures along the path are shown with B-loop depicted in green and A-loop in red. White areas were not sampled in the umbrella sampling protocol.

The two minima corresponding to the stable states of the RNA hairpin are distinct on the two-dimensional free energy map along the RMSD coordinates. When the system is in the basin encoding for the B form, the RMSD from the reference helix fluctuates around ~1-2 Å, while the RMSD from the A form is relatively high, at ~10 Å. Similar scenario is observed when the system is in the basin encoding for the A conformation. Mechanistically, the first step in the transition is the unfolding of the B-hairpin. This process starts with the breaking of the terminal base pair furthest from the loop (G5-C20), followed by partial loss of helicity and the consecutive opening of the base pairs toward the loop region (Fig. 5.14).

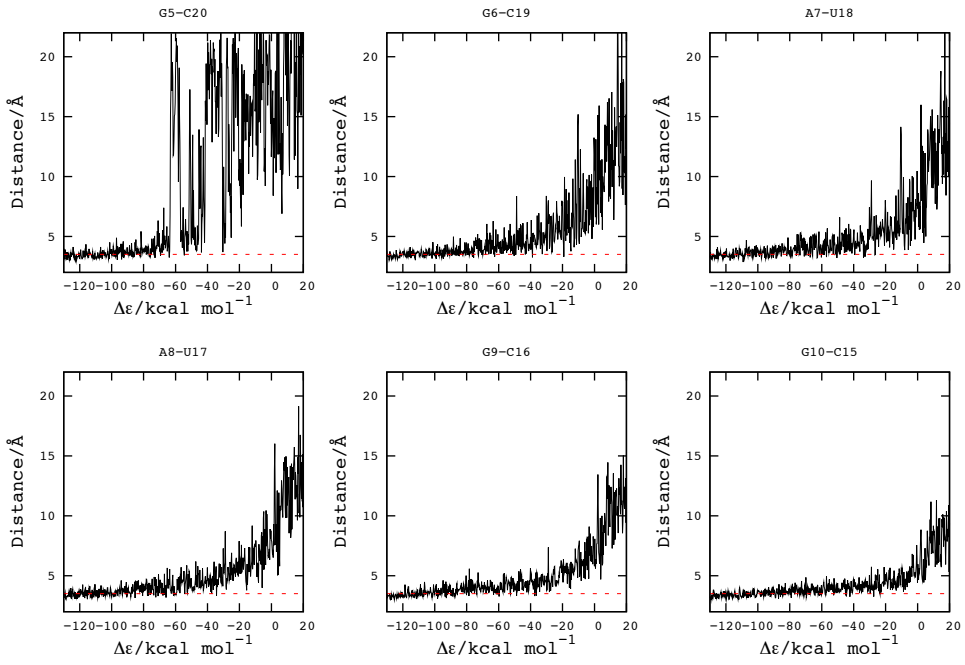

Figure 5.14 The distance between the centres of mass of the heavy atoms involved in the native base pairs is plotted as a function of the energy gap reaction coordinate in pathway 1. Umbrella windows between -120 kcal mol^{-1} and 20 kcal mol^{-1} are used to show the opening mechanism of the B helix. Horizontal red line shows the maximum value of 3.5 Å for an intact base pair.

In detail, after the fraying of the terminal base pair G5-C20 (-60 kcal mol^{-1} < Δε <-40 kcal mol^{-1}), the partial detachment of the three successive base pairs is observed (G6-C20, A7-U18, A8-U17), as shown by an increase in the centre of mass distance from the initial value ~3.3 ± 0.2 Å to ~5.1 ± 1.2 Å. Slightly more stable are the bases closer

to the loop (G9-C16 and G10-C15), which undergo a later increase in distance and less fluctuations (~4.5 ± 0.6 Å). Finally, the detachment of the two strands of the helix takes place at this reaction coordinate, between -20 kcal mol^{-1} < Δε < 0 kcal mol^{-1}. At this stage, the tetraloop (5'-UCCG-3') appears to prevent the system from the complete unfolding toward the single strand, as shown by the RMSD change along the reaction coordinate (Fig. 5.15a). When the helix is already open, the RMSD calculated with respect to the B loop is still stable around ~1 Å (Fig. 5.15a).

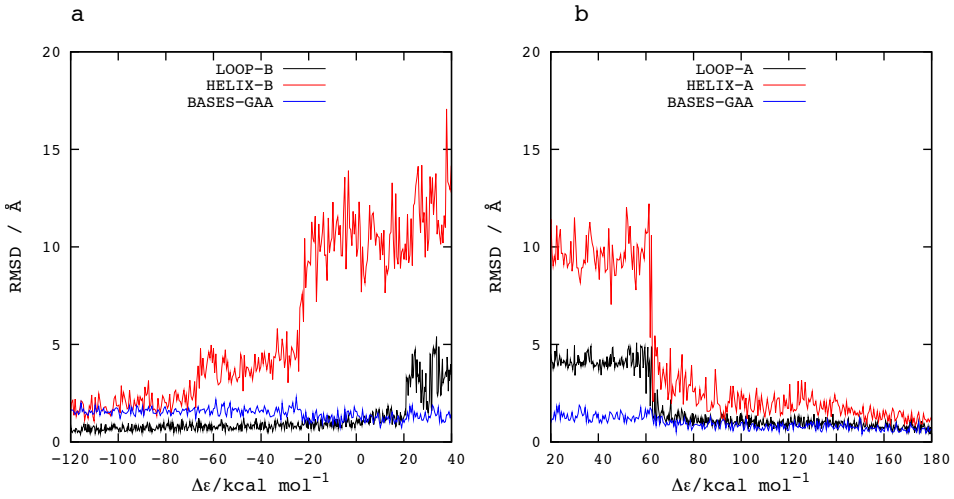

Figure 5.15 The all atom RMSD calculated with respect to the loops and helices for states B and A as a function of the energy gap reaction coordinate in pathway 1 to show the unfolding of B form (a) and the refolding of A form (b).

As a result, the structure is slightly bent around the loop, although no base pair is present. Finally, the unfolding of the tetra loop leads the system to the transition state region (35 kcal mol^{-1} < Δε < 65 kcal mol^{-1}). As shown in Fig. 5.13, the transition state region (TS_1) exhibit an RMSD larger than 11 Å with respect to both helices ($RMSD_{HELIX-A}$=11.3±2.3 Å; $RMSD_{HELIX-B}$=13.0±1.8 Å), indicating a highly unstructured state. Visual analysis of the transition state ensemble reveals an unfolded single strand in which no base pairs are present. A more detailed analysis on the transition state ensemble will be provided below, based on P_{fold} analysis.

The refolding process toward the A form is initiated by the A loop formation and subsequently zipping of the base pairs along the helix A, in a reverse mechanistic order compared to the transition from state B to TS_1 (Fig. 5.15b). The A form loop 5'-GGAA-3' is very stable due to the high degree of stacking interaction between the residues

GAA. As demonstrated by thermodynamic studies [331], this characteristic compact organization has also been observed in other crystal structures, mainly from SRP ribonucleoprotein [325, 332-334]. Consequently the GAA bases remain stacked along the entire trajectory (Fig. 5.15), although with higher fluctuations in the transition region. The closing of the first base pair G5-A8 forces the single strand to bend toward the A form.

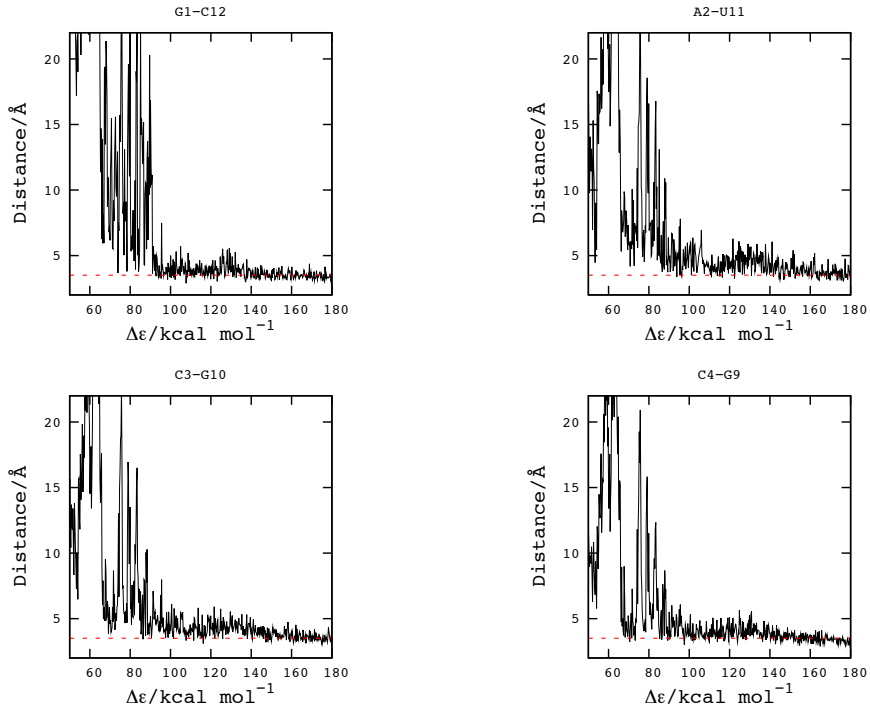

Figure 5.16 The distance between the centres of mass of the heavy atoms involved in the native base pairs of form A is plotted as a function of the energy gap reaction coordinate in pathway 1. Umbrella windows between 50 kcal mol^{-1} and 180 kcal mol^{-1} are used to show the closing mechanism of the A helix. Horizontal red line shows the maximum value of 3.5 Å for an intact base pair.

After the complete folding of the A loop the two strands are in proximity as shown by the distance of the centres of mass of relevant base pairs (Fig. 5.16), fluctuating around 4 Å (90 kcal mol^{-1} < Δε < 140 kcal mol^{-1}). In this region, the system oscillates between an open form and a partially closed helix where base pairs are occasionally formed. By the energy gap value of Δε ~160 kcal mol^{-1}, all the native base pairs form and the system stabilises in the A form, as shown by the reduced fluctuations around the reference value.

An alternative pathway has also been observed, which involves a different transition state ensemble, herein called TS_2. In this region, unlike for TS_1 ensemble, the structural similarity to both helix A and B is higher as shown by the low RMSD values (Fig. 5.17).

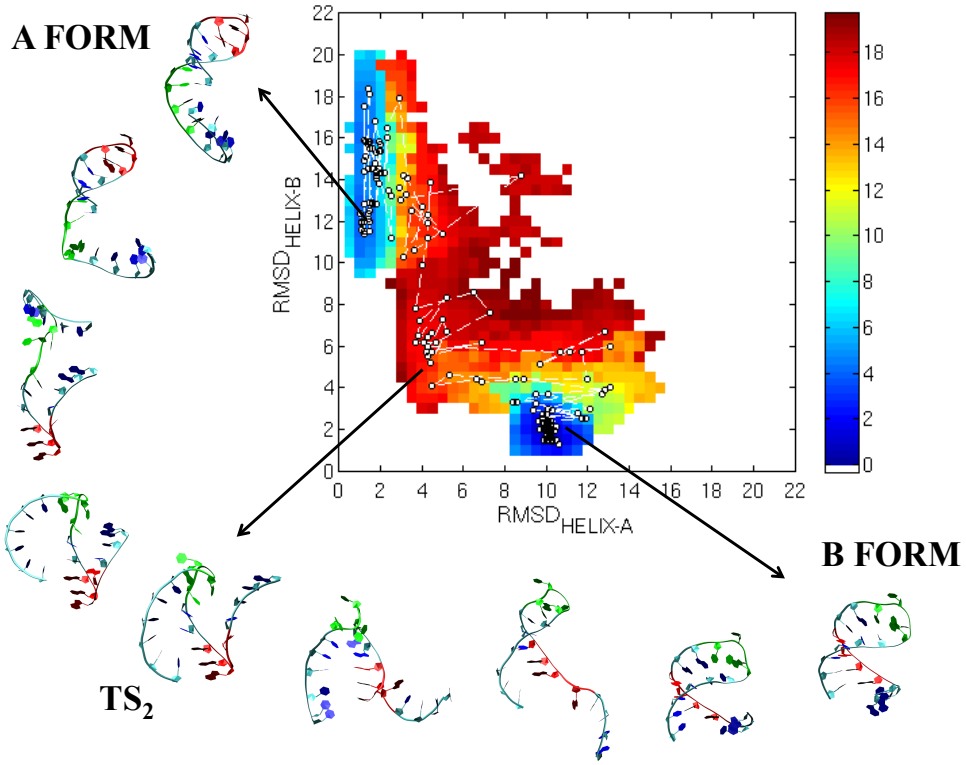

Figure 5.17 The free energy surface sampled by an individual trajectory started in form B (pathway 2) is shown as a function of the RMSD from the A form helix and from the B form helix. The white line shows the approximate path traversed during the simulation and the white points are obtained by averaging RMSD values in each window. Representative structures along the path are shown with B-loop depicted in green and A-loop in red. White areas were not sampled in the umbrella sampling protocol.

Similarly to the first case discussed above, after an initial exploration of the energy landscape in the region surrounding the basin corresponding to the B state, the helix opening represents the main structural change toward the transition state region. The unzipping starts again with the fraying of the terminal base pair (G5-C20) and is followed by a concerted opening of the helix strands, allowing the simultaneous detachment of the bases closer to the loop region. The main difference in terms of conformational changes is however observed in the sequence of events, which characterise the unfolding/folding of the two loops. In the first scenario discussed above, the unfolding of the B loop preceded the formation of the A loop, generating a

transition state region (TS_1) in which both loops are completely unfolded. In contrast, in the current pathway both loop A and B coexist in the TS_2 region (Fig. 5.18).

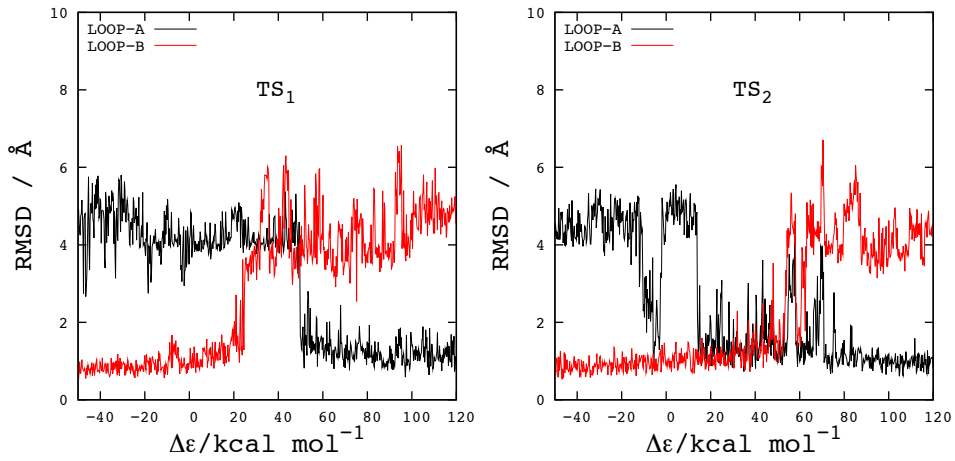

Figure 5.18 Structural comparison of the transition state region of pathway 1 (left) and pathway 2 (right). The all-atom RMSD calculated with respect to the residues forming the B loop (5'-UCCG-3') and the A loop (5'-GGAA-3') plotted along the energy gap reaction coordinate.

Structurally, the simultaneous existence of the two loops results in the alignment of the backbone characteristic of both helices A and B, and thus resulting in low RMSD values in the TS_2 region (Fig. 5.18). Visual analysis of the structural ensemble clearly shows a compact S-shaped structure, reminiscent of a pseudoknot but without the actual stabilisation provided by the native Watson-Crick base pairs. The system relaxes from TS_2 to the A form by unfolding loop B and hence rendering bases U11 and C12 available to form the terminal base pairs of helix A. At this stage ($\Delta\varepsilon$>70 kcal mol^{-1}) all native base pairs of helix A can form (Fig. 5.17).

The third example shows a transition pathway between RNA hairpin forms A and B, which traverses both transition state regions, TS_1 and TS_2 (Fig. 5.19), combining aspects of the pathways already discussed above. Thus, in this case, transition takes place via an interconversion between the compact transition structure (TS_2) and the disordered, single-stranded transition state (TS_1).

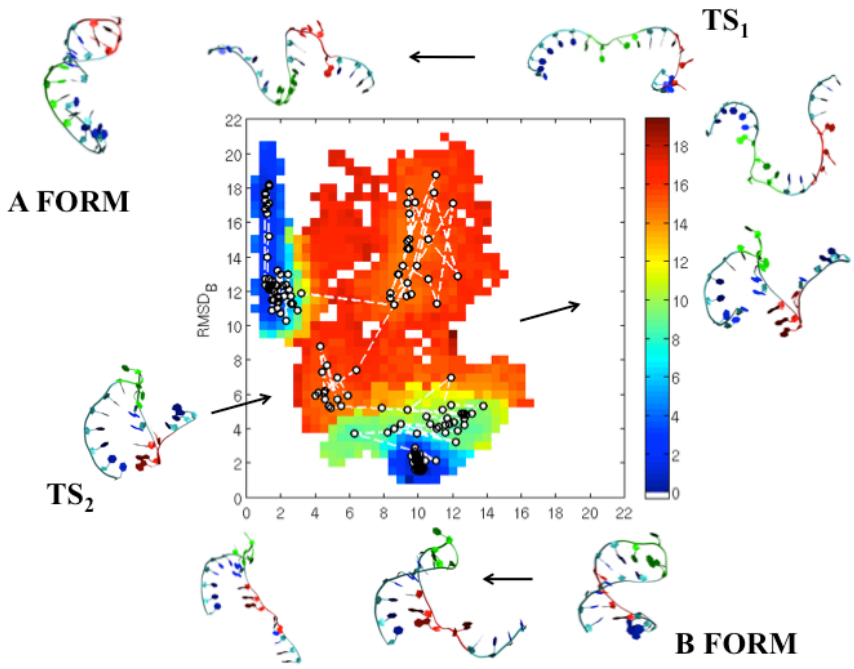

Figure 5.19 The free energy surface sampled by an individual trajectory started in form B (pathway 3) is shown as a function of the RMSD from the A form helix and from the B-form helix. The white line shows the approximate path traversed during the simulation and the white points are obtained by averaging RMSD values in each window. Representative structures along the path are shown with B-loop depicted in green and A-loop in red. White areas were not sampled in the umbrella sampling protocol.

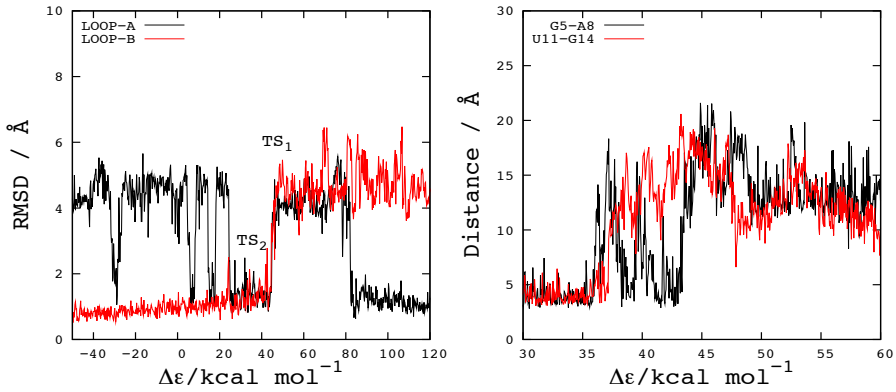

Figure 5.20 Structural comparisons of the transition state region in pathway 3. The all-atom RMSD calculated with respect to the residues forming the B loop (5'-UCCG-3') and the A loop (5'-GGAA-3') plotted along the energy gap reaction coordinate. (left panel). Opening of the capping bases of the two tetraloops shown by plotting the distance between U11-O2 and G14-N1 (loop B) and between G5-N2 and A8-N7 (loopA) (right panel).

The interconversion of the transition structures is clearly shown by the all-atom RMSD calculated with respect to loops A and B (Fig. 5.20). The opening of the loops is driven by the disruption of the capping base pairs U11-G14 (B-form) and G5-A8 (A-form). In particular, the analysis of the interatomic distances between U11-O2 and G14-N1 and between G5-N2 and A8-N7 (Fig. 5.20) shows that loop A oscillates between closed and

open states before the final opening transition to TS_1. The existence of such a complex transition pathway shows that the general energy gap reaction coordinate does not directly bias the structural details of the transition and the two competing pathways can interconvert. Putting together the data from the 50 simulated trajectories, the results show that 65% pass through pathway 1 via TS_1, 10% pass through pathway 2 via TS_2, and the remaining 25% of the trajectories involve transitions between TS_1 and TS_2 as described above in the third example.

5.8 Pseudoknot as the transition state structure

Following previous analysis, one obvious question is why in the actual model the transition through a pseudoknot structure, as suggested in the hypothesis of Wenter et al. [305] and from the Kinefold prediction (Fig. 5.21), was not observed.

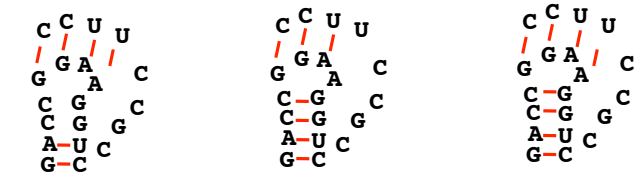

Figure 5.21 Schematic representations of pseudoknot structures as proposed by Wenter et al. [305] (6 base pairs), from Kinefold analysis (7 base pairs) and the pseudoknot model used in this work (8 base pairs).

In the following, I describe the procedure used to drive the interconversion pathway between forms A and B through a pseudoknot structure and evaluate the corresponding free energy. The first step consists of building a three-dimensional model of the pseudoknot where base pairs of helices in A and B forms coexist. For this purpose, I used Nucleic Acid Builder, part of the Amber Tools [335], where base stacking and pairing restraints can be imposed based on a distance geometry algorithm. Cycles of simulated annealing procedure provided a low-energy pseudoknot structure stabilised by 8 base pairs: 4 base pairs from helix A and 4 base pairs from helix B (Fig. 5.22).

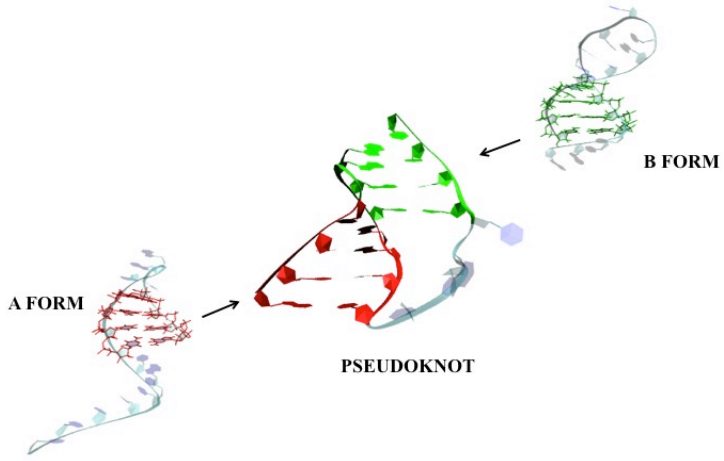

Figure 5.22 The three-dimensional model of pseudoknot (central structure). In red and green are represented, respectively, residues originating from helix A and helix B. In addition, the original A and B forms of RNA are represented on the left and right.

Since the energy gap reaction coordinate cannot enforce a particular structure to be sampled along the pathway, geometric reaction coordinates must be used in conjunction with the EVB-SBP approach. Here I use the root mean square distance (after least squares fitting) of base heavy atoms, as the reaction coordinate, between the actual atomic positions and a suitable "target". Four simulations were performed driving the system from A or B form of the RNA to the "target" pseudoknot structure, and for the reverse process, two simulations from the pseudoknot to either the A or B form RNA structure. Starting from an initial conformation, the system was restrained by RMSD in decreasing steps of 0.1 Å toward the target with a harmonic force constant of 2 kcal mol^{-1} $Å^{-2}$. At each step the value of the reaction coordinate was collected for 2ns.

Although the RMSD proves to be a suitable reaction coordinate to sample the pseudoknot structure, the associated free energy changes would show an apparent increase as the reaction coordinate goes to zero [336]. This is due to the fact that since the target is a unique structure, the 'volume' of the initially large conformational space becomes essentially zero as the system approaches the target structure. For non-linear transformation of the three-dimensional atomic coordinates to one-dimensional reaction coordinate in free energy calculations, a more involved Jacobian correction must be applied [246, 337]. Another possibility is to re-map the free energy along a reaction coordinate that does not involve such correction. Here I have calculated the potential of mean force as a function of the energy gap reaction coordinate, using the

conformational ensemble sampled along the RMSD reaction coordinate. In order to extract this information a two-dimensional WHAM procedure was used [215, 281].

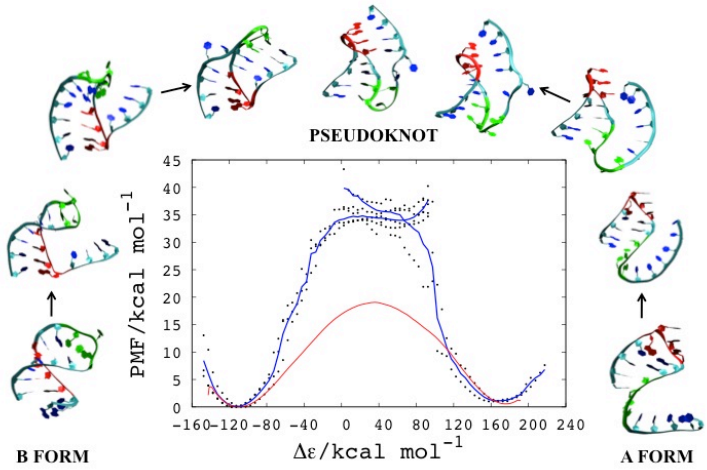

Figure 5.23 Free energy changes of the interconversion of A and B form RNA via the pseudoknot transition structure. Individual trajectories biased by the RMSD reaction coordinate are represented in dotted lines and the corresponding average free energy profile is shown in blue. For comparison, the free energy profile corresponding to the simulations biased by the energy gap reaction coordinate is shown in red.

The transition from the B form to the pseudoknot involves the disruption of only two base pairs and the formation of four new base pairs, corresponding to the helix A; the transition from the A form to pseudoknot involves the formation of 4 new base pairs, part of the helix B. Although, mechanistically only limited changes are required for the formation of the pseudoknot transition structure from the stable RNA forms, the results clearly indicate that the pseudoknot is significantly higher in free energy with respect to the unfolded transition state ensemble, according to our EVB-SBP model (Fig. 5.23).

It is worth noting that the PK structure is an extremely compact structure as evidenced by the radius of gyration R_g ~11.6 Å, which is lower than both B (R_g~11.9 Å) and A (R_g~13.4 Å) forms. Such a compact structure is highly entropically unfavourable. As previously shown (Fig. 5.10), the entropy contribution at the transition state region can be estimated by the comparison between the EVB ground state energy and the free energy. In this case, the observed activation enthalpy is $\Delta H^{\ddagger}_{B\rightarrow A}$~50 kcal mol^{-1} and the activation free energy is $\Delta G^{\ddagger}_{B\rightarrow A}$=35 kcal mol^{-1}. As a result, the entropic term in the transition state amounts to ~15 kcal mol^{-1}. It is interesting to note that the observed

entropy for a compact PK structure corresponds to only 36% of the entropy estimate (~42 kcal mol^{-1}) observed in case of a single strand transition state (Fig. 5.1). As a result, one could speculate that the high free energy of the PK structure is mainly dependent upon the limited conformational entropy of the state.

This result does not exclude the hypothesis that an associative mechanism may exist, but it provides an explanation for the absence of a pseudoknot structure among the transition structures sampled along the energy gap reaction coordinate.

The pseudoknot structure was expected to be stabilised by enhanced base pairing, compared with the single-stranded transition structure where only base stacking interactions are present. As discussed in Chapter 4 (Section 4.2), the energy function is based on simple distance cut-off in a reference structure, and as a result, the number of native contacts between stacking bases is higher than between pairing bases. Although stacking is expected to be more favourable than pairing between two bases in water, the pseudoknot may become less penalised with more favourable base pairing energies. It is possible that the simple, distance-based native energy function would need to be modified for a more accurate representation of the energetic balance in the system. This, however, would also introduce *a priori* bias whether certain bases are paired or stacked along the entire simulated pathway.

5.8 Characterisation of the transition state ensemble

In order to confirm the validity of a putative transition state ensemble, unrestrained downhill trajectories were initiated from the highest free energy structures using random initial velocities. This analysis is independent of the reaction coordinate and enables the calculation of P_{fold} values [338]. A true transition state is defined as the region in the conformational space with equal probability to access both free energy minima (P_{fold}=0.5).

For this analysis, 50 structures were selected from the putative transition state region, from each window with the highest free energy in the umbrella sampling simulations. For each structure, 50 independent unrestrained simulations, each 10 ns long, were performed, amounting to ~25 μs simulation time. To test the convergence of the P_{fold} values, data were used from increasing number of simulations, 10, 20, 30, 40, and 50. No significant difference was observed in P_{fold} values using more than 30 unrestrained simulations, as shown by fluctuations of P_{fold} values on the order of ±0.02.

As described by Hubner et al. [339], the P_{fold} calculation amounts to a Bernoulli trial, and the relative error resulting from using N runs scales as $N^{-1/2}$. Hence the 50 trial simulations estimate P_{fold} values within 14% of the mean.

Each trajectory was then classified whether the final structure has reached basin B or basin A. For this committor analysis, the energy gap value corresponding to the final structure was used: state B was considered if $\Delta\varepsilon < -60$ kcal mol^{-1} and state A if $\Delta\varepsilon > 110$ kcal mol^{-1}. Out of 2,500 simulated trajectories, about 60% folded into either A or B forms within 5 ns, 20% within 10 ns, while the remaining ~20% did not reach the defined boundaries of the stable basins within the allocated simulation time. On the whole, about 60% of the trajectories folded in A form and 40% in B form, however, with large differences among the starting structures.

To visualise the results for each starting conformation, structures were mapped onto a two-dimensional map according to their RMSD value with respect to helix A and helix B in the reference structures (Fig. 5.24). The P_{fold} value of zero (blue) means that all the trajectories from a particular starting structure end up in basin A, while 1 (red) indicate that all the trajectories from a particular starting structure end up in basin B. Structures are grouped in two main clusters: the first with low similarity (high RMSD) with respect to both helices, corresponding to unfolded structures (Fig. 5.24a right circle) and the second corresponding to a small group of compact structures with higher similarity to helix A and B (Fig. 5.24a left circle). In both clusters P_{fold} values are ranging from 0 to 1, with some structures corresponding to 0.5. In order to obtain a more detailed picture, P_{fold} values for the same structures are also shown as a function of the RMSD calculated with respect to loop A and loop B. In this case three subgroups can be indentified: i) structures where both loops are folded (Fig. 2.24b left circle); ii) structures where only loop B is formed (Fig. 5.24b bottom right circle); iii) structures where both loops are unfolded (Fig. 5.24b top right circle).

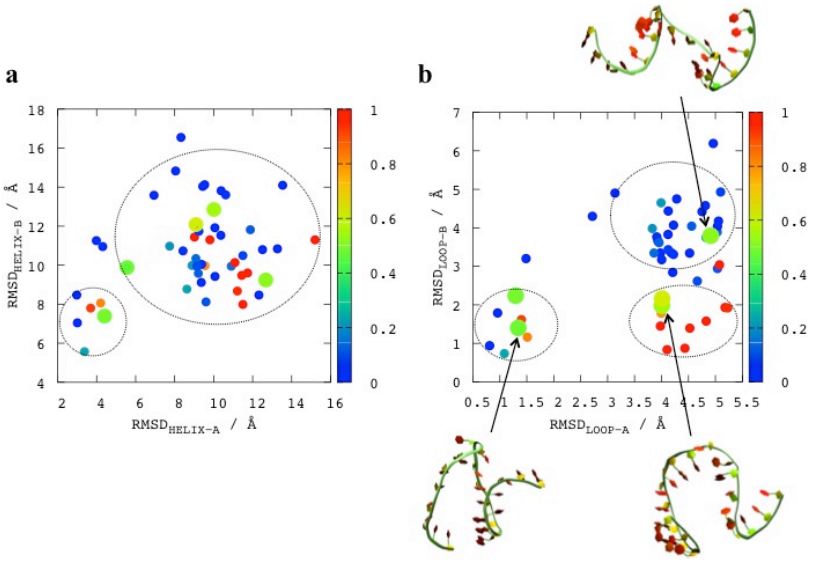

Figure 5.24 P_{fold} values for each structure are shown as a function of RMSD calculated with respect to A and B forms using helix residues (a) and loop residues (b). Circles indicate subgroups of P_{fold} values. Selected structures corresponding to P_{fold} values ~0.5 are shown.

The results indicate that folding toward the A or B form is strongly dependent upon loop formation. When both loops are formed then structures can fold toward either B or A form. It has to be noted that within this subgroup (Fig. 5.24b left circle), similar structures reveal diverse P_{fold} values. Structures where RMSD with respect to loop A is <1 Å fold in A form (P_{fold} ~0), while structures where RMSD with respect to loop A is >1 Å fold preferentially toward the B form (P_{fold} ~1). A visual analysis reveals that in the first case the loop closing base pair G5-A8 is perfectly formed, while in the second case it is not formed with G5 pointing out of the loop. If only loop B is folded (Fig. 5.24b bottom right circle), the preferred route is toward the B form, as indicated by P_{fold} values ~1. When both loops are unfolded (Fig. 5.24b top right circle), the preferred folding path is toward the A form as evidenced by P_{fold} values ~0. This could be explained by the fact that in the actual parameterisation of the model, the B loop formation, although a higher scaling factor for native interactions was used (see above), is still slightly less favourable than the A loop formation.

Within the whole ensemble, only few genuine transition state structures were identified with P_{fold} ~0.5. An analysis of these structures shows that the radius of gyration ranges from 13.5 to 15.5 Å, depending on the conformational arrangement of the single-stranded part of the RNA. Interestingly, no native hydrogen bonds were observed in these structures with distance of donor and acceptor <3.5 Å. This provides further evidence that the transition state is characterized by the formation of an

unstructured single strand, where the formation of native hydrogen bonds is not a relevant feature of the transition mechanism. The role of loop stability in defining the transition structure became apparent, in line with experimental finding that tetraloops initiate hairpin folding.

5.9 NMR spectroscopy

I carried out NMR experiments on the 20nt RNA sequence, in collaboration with Jean-Louis Leroy (ICSN, Gif-sur-Yvette, CNRS, France), to provide further insight into the mechanism of switching of the bistable RNA and compare those with the theoretical findings discussed above.

The first part of this experimental work was focused on the study of base pair opening lifetimes. Studies pioneered by Gueron et al. [252] described how NMR imino proton exchange experiments, can be applied to unravel the base pair opening process. Imino proton exchange happens through base pair opening followed by the exchange of the open state. The net exchange rate is therefore distinct from that of the base pair opening, unless exchange occurs at every opening. In order to increase the exchange rate, it is possible to use imino proton exchange catalysts (i.e. chemical bases such as NH_3). As a result one can extrapolate the "true" base pair lifetime at infinite catalyst concentration, assuming that the exchange happens at each opening event. Additionally in the second part of this work, NMR experiments were performed in order to study the ratio of A/B conformations at equilibrium and far from equilibrium. The aim of this study was to gain insight into the nature of the transition structure between fold A and fold B.

I note that previous NMR spectroscopy experiments were performed on the same system, see Section 5.3.1 and ref. [305]. Residue-specific imino proton exchange times are measured for the bistable RNA system and compared to published data.

5.9.1 Experimental methods

Oligoribonucleotide Synthesis and Sample Preparation: The oligoribonucleotides were synthesized with cyanoethyl phosphoramidites on a 2 μM scale. These were then purified on a Q sepharose Hiload column using a NaCl gradient. After purification, the oligomers were dialyzed against water and lyophilized. The NMR samples were

prepared by dissolving the oligoribonucleotides in a 90% H_2O, 10% D_2O solution containing 0.1 M NaCl, 1 mM ethylenediamine tetraacetic acid and 0.2 mM 2,2-dimethyl-2-silapentane-5-sulfonate. For the imino proton exchange experiments catalyst was added to the NMR sample from 0.6 M ammonia pH 8.8 stock solutions.

Aliquots of the NMR sample were analysed using gel filtration chromatography. Systematic addition of thymidine at a micromolar concentration to the oligo provided a reference marker on the chromatograms. The elution buffer was 0.4 NaCl and the elution time ranged from 5 to 8 minutes.

NMR spectroscopic methods: 1D H NMR spectra were collected using a 500 MHz Varian INOVA spectrometer. Imino proton exchange measurements were performed using water magnetisation transfer. Selective inversion of the water magnetisation was realised through DANTE sequence [340]. The magnetisation of the imino proton was detected after a variable delay, incremented from 1 to 210 ms using an echo sequence as previously described [341]. Spurious effects due to cross-relaxation were limited by using variable delays shorter than 200 ms. The exchange contribution of the added catalyst was determined from the exchange times measured in the presence ($\tau_{ex,cat}$) and absence (τ_{ex}) of the catalyst.

5.9.2 Experimental results

An aliquot of the NMR sample was analyzed through gel filtration chromatography (Fig. 5.25a) to evaluate the purity of the sample and the propensity to form duplex segments from the two complementary single stranded oligomers of the 20nt sequence. The equilibrated sample shows the presence of ~10% dimeric species. However, repeating the analysis after melting and fast cooling, a fully monomeric sample is obtained.

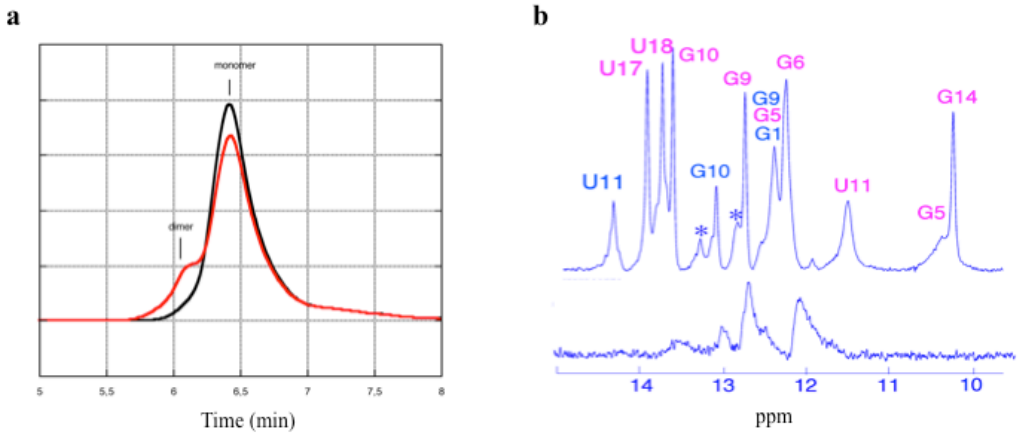

Figure 5.25 a) Gel filtration chromatogram of the 20nt RNA at temperature 298 K (red line) and after melting and fast cooling to 273 K (black line); b) 1D H-NMR spectrum at 288 K and pH 6.3 (top spectrum) and after addition of 0.6M ammonia solution pH 8.8 at 288 K.

Imino proton spectra were collected by 1D H-NMR at T=288 K and pH 6.3 (Fig. 5.25b) and assigned using data from the literature [305]. The comparison between the spectra with and without added catalyst shows significant line broadening effects due to the exchange of imino protons with the surrounding solvent. In theory, the increase in exchange rates, observed in the presence of catalyst, correspond also to an equal increase in the relaxation rates, and hence relaxation measurements could be applied to measure exchange rates [82]. We used magnetisation transfer to measure imino proton exchange rates. The longitudinal magnetisation of the imino proton is measured after a variable delay ranging from 1 to 210 ms, following the selective inversion of the water magnetisation (Fig. 5.26).

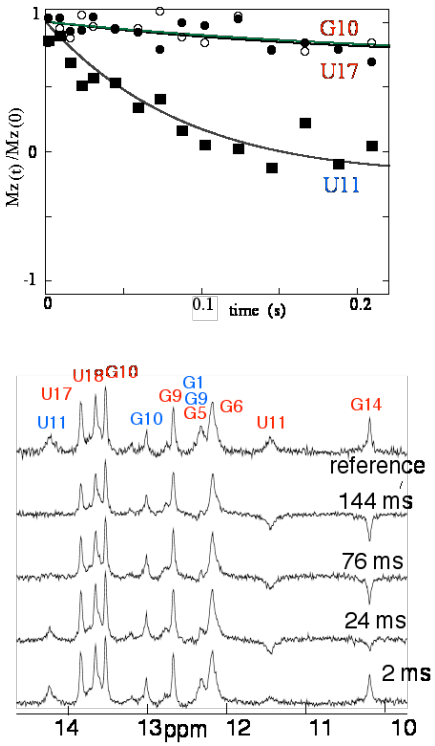

Figure 5.26 Imino proton exchange time measurements. Top: The longitudinal magnetisation of selected imino protons was measured after variable time delay between selective inversion of the water magnetization and the detection pulse. The lines are computed according to Gueron et al. [82] and give the imino proton exchange time values displayed in Table 5.4. Bottom: Illustration of the magnetisation transferred from water to the imino proton after the variable delays indicated on the assigned spectra.

Exchange times are calculated by fitting τ_{ex} as the adjustable parameter to the decay rate of the magnetisation as a function of time following the procedure described earlier [82]. In addition, the effect of added catalyst on the exchange time was evaluated ($\tau_{ex,cat}$). The calculated imino proton exchange times are summarised in Table. 5.4 along with published values [305] for comparison purposes.

Table 5.4 Imino proton exchange times (units in sec) measured at 288 K.

		Fold B				Fold A		
		τ_{ex} (a)	τ_{ex} (b)	$\tau_{ex,cat}$(c)		τ_{ex} (a)	τ_{ex}(b)	$\tau_{ex,cat}$(c)
Loop	U11	0.006	1.326	<0.002				
	G14	0.016		<0.002				
Stem	G10•C15	0.79	<0.002	<0.002	C4•G9	<0.05	0.463	<0.002
	G9•C16	1.04	0.023	0.038	C3•G10	0.72	0.015	0.0075
	A8•U17	0.714	0.222	<0.002	A2•U11	0.138	0.878	<0.002
	A7•U18	1.21	0.084	<0.002	G1•C12	< 0.05	0.711	<0.002
	G6•C19	1.52	0.022	0.177				
	G5•C20	<0.05	n.d.	<0.002				

(a) without catalyst at pH 6.3; (b) without catalyst, but derived from the intensity of cross peaks observed between the imino proton, water and the neighbouring protons from Wenter et al. [305] for comparison purposes; (c) in the presence of catalyst: 0.6 M ammonia solution at pH 8.8.

First, it can be seen that the imino proton exchange times measured with magnetisation transfer differ significantly from those deduced from 2D NOESY crosspeaks of imino protons and water [342]. The method of magnetization transfer has been used and cross-validated successfully in the past. It is important to emphasise, however, that imino proton exchange times involve the net effect of conformational change to expose the exchangeable site and the actual chemical process of proton exchange. In the absence of added catalyst, there is no way of knowing if the exchange time is related to conformational changes. By adding increasing amount of catalyst and extrapolating to infinite catalyst concentration, the exchange time shortens and ultimately the true base pair lifetime can be obtained. Here, the introduction of an external catalyst accelerates the imino proton exchange rates, with major effects on the helix stem terminal base pairs and on the A-U Watson Crick pairs. This is also evidenced by broadening of the corresponding NMR signals, as shown in the NMR spectrum (Fig. 5.26). The comparison between exchange times in the presence and absence of added catalyst clearly shows that the exchange rates without catalyst are not related to the base pair lifetime and hence to the actual conformational change.

Bulk kinetic experiments require a starting condition which keeps the system far from equilibrium before measurements commence. Wenter et al. achieved this in an elegant experiment by introducing a photolabile group that preferentially stabilised fold A. Since equilibrium NMR spectroscopy is not capable of measuring properties of transient structures, we have planned to carry out an "experimental P_{fold}" analysis,

normally performed using computational methods [338], where the system is kept in the transition state region, and subsequently allowed to relax to one of the stable structures. The first measurement time should be chosen such that downhill folding can take place but not the interconversion between the stable states. In such a scenario, it is expected that a ratio of 50:50 for the two stable states is obtained, before interconversion can take place. It is important to stress that this hypothesis is only valid assuming the free energy surface is symmetric around the transition state region, hence equal curvature of the potential both toward the A and B forms. However, in case of a non-symmetric free energy surface a different probability of folding toward the A and B forms could be observed and the original assumption may not be true.

To test the possibility of a single-stranded transition state for the interconversion process, the 20nt RNA oligomer was heated up to ~373 K in the NMR tube to obtain completely unfolded (melted) oligomer. The sample was subsequently put on ice and NMR spectra were recorded after 1 min, 10 min, 1h, 2h, 6h at 288 K to follow the potential equilibration process.

Based on the temperature dependence of the interconversion rate constant [305], the rate at 273 K can be obtained by extrapolation (Fig. 5.27). Note that the experimental interconversion rate at 273 K or 288 K [305] is significantly longer than the time used to record the first NMR spectrum. It is remarked that during the cooling process from 373 to 273 K, the system traverses elevated temperatures, where interconversion rates are still fast (eg., at 323 K, the extrapolated lifetime is about 0.1 s). This means that if the cooling is not fast enough, the system may not only fold to one of the stable forms but also reach equilibrium between form A and B before reaching 273 K. It should be noted that extrapolated rate constants are not relevant at temperatures above (and probably close to) the melting temperature of 351 K [304] where only the unfolded single strand is stable and the extrapolated rate constants are equal.

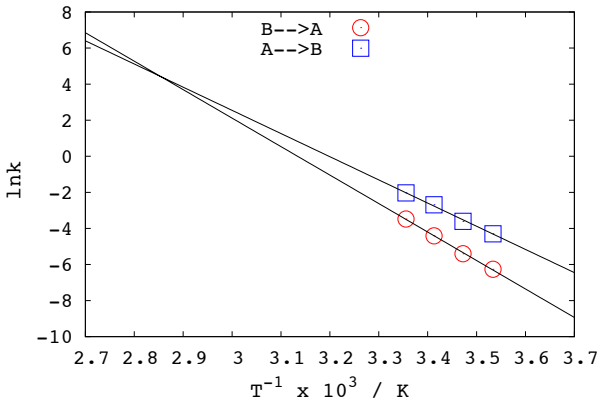

Figure 5.27 Arrhenius plot of experimental rate constants [305] for A→B and B→A transitions.

The NMR spectra of the quenched sample were recorded at 288 K at different time intervals after quenching (Fig. 5.28).

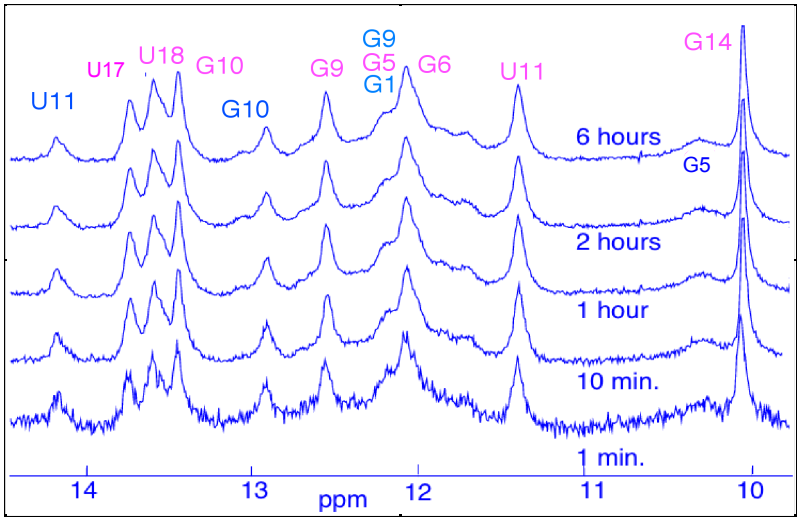

Figure 5.28 Imino proton region of H-NMR spectra recorded at various delay times from snap cooling at 288K.

In the first spectrum, recorded after snap cooling, peaks corresponding to both forms A and B are visible. An integration of the relevant peaks provide a ratio of A/B forms of 28:72, very close to previous results of equilibrium measurement of 25:75 [304]. This would suggest that the cooling of the sample was not fast enough to prevent interconversion and equilibration. One could propose that using a different experimental setup, where faster cooling can be achieved, the initial ratio of 50:50 would equilibrate

to 25:75 over time. Nevertheless, this experiment suggests that the single-stranded RNA structure can directly access forms A and B, and hence it is a potential transition structure in the interconversion pathway.

Putting together all experimental findings, the hypothesis of an associative mechanism proposed by Wenter et al. is not supported [305]. As described above, this hypothesis is based on the assertion that the activation enthalpies ($\Delta H^{\ddagger}_{(A\rightarrow B)}$=25.5 kcal mol^{-1}; $\Delta H^{\ddagger}_{(B\rightarrow A)}$=30.6 kcal mol^{-1}) amounts to about half of the entire base pair enthalpy determined by thermal denaturation studies of reference systems [304]. In this case, a full unfolding of the hairpins should not be possible. The reference systems involve truncated sequences with only one possible hairpin conformation each: 5'-GACCGGAAGGUCC-3' corresponds to fold A, and 5'-CGGAAGGUCCGCCUUCC-3' to fold B. It is noteworthy, however, that the truncated sequence for fold A is more stable (ΔG=-7.9 kcal mol^{-1}) than the one for fold B (ΔG=-7.3 kcal mol^{-1}) [304], in contrast with the stabilities observed for the full system. Therefore, the use of truncated sequences as a reference may not be appropriate. Nevertheless, if we consider the free energy of activation (Table 5.3), instead of the enthalpy of activation, the denaturation of the truncated systems require only ~7.6 kcal mol^{-1}. This value is significantly below the activation free energy of the interconversion process and thus the complete unfolding of the hairpin forms should be possible. This analysis shows that the free energy changes of the process suggests a dissociative mechanism as a possibility.

Furthermore, the positive activation entropy calculated here (Table 5.3) appears to contradict an associative mechanism involving a compact transition state. In addition, it was suggested that the hairpin unfolding is initiated by the disruption of base pairs close to the loop region in an associative process [305], based on the comparison of imino proton exchange rates of the 20nt RNA with those of the truncated sequences. However, as shown above, imino proton exchange rates do not report on conformational changes without added catalyst. Several recent studies have demonstrated [200, 343-346], that the dominant pathway of hairpin unfolding involves an unzipping mechanism. This process starts with the fraying of the terminal base pair at the end of the helix stem and not from the region close to the loop [344]. Although this experimental study cannot exclude the possibility of an associative mechanism, all results suggest a dissociative mechanism with a single-stranded transition state ensemble, supporting the theoretical findings.

5.10 Summary

The switching mechanism of a model system, a bistable RNA sequence, was studied using the EVB-SBP approach in combination with solution NMR measurements. It is important to understand the mechanism at the atomic level to aid the studies of more complex biological machineries, such as the RNA riboswitches. Using previous experimental information, the free energy surface was parameterised according to the EVB procedure. An extensive amount of simulations was performed including both restrained and unrestrained simulations, totalling ~42 μs. It is found that the short RNA sequence exhibits heterogeneity in the interconversion mechanism, with two main competing pathways identified. One pathway involves an unfolded RNA strand, while the other pathway involves a more compact transition structure, but without stabilising Watson-Crick hydrogen bonds expected for a pseudoknot. The energy gap has proven to be a suitable reaction coordinate for complex systems, which does not bias the structural details of the mechanism. Characterisation of high-free energy structures through P_{fold} analysis further delineated the transition structure. Experimental support was provided through NMR spectroscopy, which are in line with our theoretical model. It has shown that imino proton exchange measurements performed without added catalyst cannot report on conformational change or base pair lifetimes. A new "experimental P_{fold}" analysis was designed to better understand the transition state ensemble. However, the interconversion rate between the stable states appeared to be faster than the quenching time.

Chapter 6

Concluding remarks and future work

Biological macromolecules are the “small workers” which give shape to life. They play a key role in every biological process, demonstrated by the extremely wide range of known functions. Despite their enormous diversity, a common feature of biological molecules is the strict relation between sequence, structure and function. Nevertheless, biomolecules are highly flexible, hence understanding their dynamics may be essential in order to bridge the gap between structure and function. Due to the complexity of these processes, it remains challenging to describe the dynamics at the atomic level for both theoretical and experimental approaches. In particular, computational techniques have been largely applied to study these processes at the atomic level. However, time scales affordable by classical computational methods are orders of magnitude lower than time scales of relevant biological processes. In the past decades, huge efforts have been directed toward building powerful computers and developing efficient theoretical methods. The present thesis aimed at introducing a new computational tool and applying it to various biological problems. The new method, described in detail in Chapter 3, was developed with the aim to combine theory and experiments. A thorough validation of the method was carried out using a simple model system. The first application of the method was the study of base flipping in B DNA (Chapter 4). The main application of the method was the study of conformational switching mechanism of bistable RNA, as described in Chapter 5. Finally, the present work is concluded by reflecting on the new method developed and describing some recent application on a protein system, which will be part of future work.

6.1 EVB-SBP method and its applications

Structure-based potentials were used to reproduce structure and dynamics of biomolecules from reference all-atom force field simulations. The actual implementation of SBP used here is fairly simple, with only two adjustable parameters: the cut-off and the scaling factor. The cut-off establishes the number, while the scaling factor determines the energetics, of native interactions. However, one of the main limitations of SBP is that they are not transferable, i.e., they cannot be extended to different systems. Various parameterisation approaches have been tested, including fitting the potential to reproduce structural data, energy gradients and atomic distance distributions from reference simulations. The equilibrium distances in the native 12-6 LJ potential may come from the sum of the atomic van der Waals radii or directly from the reference structure. Standard structure-based potentials, however, encode for one dominant minimum, which prevents their applications to complex conformational changes, where more than one stable state is involved.

Multiple structure-based potentials were coupled by the empirical valence bond theory to provide a unified potential energy surface. Two parameters were used to adjust relative energies of the basins and the height of the energy barrier, the diabatic energy shift and the coupling constant, in order to reproduce experimental data. On the parameterised energy surface, either direct molecular dynamics or umbrella sampling simulations can be performed to sample the phase space. The conformational transitions were driven along the energy gap reaction coordinate, constituting the energy difference of diabatic states.

The EVB-SBP method has been implemented in the `sander` module of the Amber package, allowing various functional forms for the native interactions. The implementation has been thoroughly tested on a model system, and compared with analytical results. Subsequently, it has been applied to DNA, RNA and protein systems.

The first application included the study of base flipping in B-DNA. This work follows a series of earlier studies, which have used geometric reaction coordinates to elucidate the mechanism of base flipping [83, 85, 258]. In this work, it has been shown that the energy gap reaction coordinate can induce base flipping without biasing the directionality of the process. Base rotations toward the major and minor grooves were observed in an ensemble of umbrella sampling simulations. Interestingly, an alternative,

high free energy pathway has also been observed, where base closing resulted in a *syn* conformation of the base stacked in the DNA helix.

Next, the switching mechanism between two hairpin loop structures of a 20nt bistable RNA was studied. Two alternative pathways were predicted at the secondary structure level by the program Kinefold. The associative pathway involves a pseudoknot structure with intercalated helices of the two stable states. The dissociative pathway involves a single stranded oligomer. After constructing the three-dimensional models of the transition endpoints, all-atom force field simulations were carried out. The EVB-SBP approach was then employed to study the interconversion mechanism between the two endpoints. The free energy surface was adjusted to reproduce experimental data [305]. Umbrella sampling simulations in conjunction with the general energy gap reaction coordinate were employed, totalling ~16 μs of simulation time. An analysis of the transition pathways showed structural heterogeneity in the interconversion mechanism with two main pathways identified. One pathway traverses a compact transition state, where both loops are folded, but without the formation of stabilising base pairs of the pseudoknot structure. The other pathway traverses an unfolded single strand in a dissociative manner. In this case, the molecular mechanism shows unzipping of the base pairs starting from the helical end and unfolding of the hairpin loop to form the single strand, followed by formation of the new loop and zipping of the new helix in a reverse mechanistic order. Interestingly, an exchange between different pathways has also been observed. The transition state ensemble was studied using a P_{fold} analysis. Unrestrained simulations were initiated from the putative transition state region, totalling about ~25 μs. This pointed to a heterogeneous transition state ensemble, characterised by a single strand with diverse structural organizations.

In order to support the theoretical model proposed for the RNA switching mechanism, NMR experiments have also been performed as part of this thesis work. In particular, imino proton exchange rates were measured on the model system and an "experimental P_{fold}" analysis was attempted.

Experimental findings are shown to be in agreement with the theoretical model. In particular the "experimental P_{fold}" suggests that the single-stranded RNA can directly access forms A and B, and hence it is a potential transition structure in the interconversion pathway. Previously, Wenter et al. [305] suggested an associative mechanism for the interconversion pathway, which is in disagreement with the present findings. The associative mechanism, in fact, was never directly observed in the EVB-

SBP simulations, hence a geometric reaction coordinate was used to bias the system toward the pseudoknot transition structure. The theoretical results confirmed the higher free energy of this state with respect to the unfolded transition state. It should be noted that the data provided by Wenter et al. clearly indicate positive activation entropy, which is in line with a dissociative mechanism rather than an associative one. In conclusion, although it is not possible to exclude a pseudoknot as the transition state, several lines of evidence have indicated a dissociative mechanism as the dominant pathway, which governs the RNA switching mechanism.

6.2 EVB-SBP method: advantages, limitations and future developments

Structure-based potentials have been used in several previous works [194, 195, 202, 344]. The main advantage of using SBP is that it provides a funnel-shaped energy landscape, at low computational cost, where the native state is the global minimum. Usually, SBP are used in conjunction with coarse-grained representation of the molecular structure. Very recently, parallel to this work, all-atom SBP has been developed for both protein [224] and nucleic acids [226], providing, in principle, a higher degree of accuracy. In this thesis, atomistic structure-based potentials were coupled using empirical valence bond theory, which provide a parameterised multiple-basin energy surface. A similar approach has been applied for a coarse-grained protein system [203]. The EVB-SBP method can be viewed as bridging the gap between the classical all-atom simulations, accurate but computationally expensive, and the earlier simplified models which were computationally efficient but with a lower degree of accuracy.

However, some considerations should be made regarding the limitations and possible future improvements of the method. Firstly, non-native interactions in the structure-based potential were considered only repulsive. A more accurate model could consider non-native interactions either neutral or slightly attractive, instead of repulsive. Although this would introduce more frustration to the potential energy surface, it is arguably more accurate in describing regions of the conformational space far from the minimum, where non-native interactions may play a role. Alternative potentials could be used for this purpose, such as the Weeks-Chandler-Andersen [347] potential, which is not repulsive beyond the reference distance.

In addition, in the actual SBP, two parameters, a distance cut-off and an empirically chosen scaling factor, have been used. The study of the RNA system, however, has revealed that a simple distance-based cut-off may result in stacking interactions that are slightly overweight with respect to pairing interactions. In the future, one may think of using a different scaling procedure for pairing and stacking interactions, to achieve a more accurate energy balance between the two types of interactions. This would, however, necessitate the *a priori* classification of nucleotide pairs in the stable states.

Another feature of the EVB parameterisation procedure consists of tuning the diabatic state shift and coupling element to construct the combined potential energy surface. In this model, a constant coupling element has been employed, which modifies the shape of the surface at the crossing of the diabatic potentials. However, a non-constant coupling element, such as exponential or Gaussian functions of the coordinates may be useful for a better representation of the transition state region.

Improvements could also be made in order to make the code more computationally efficient and fast. A modification would consist of simplifying the procedure of calculating repulsive non-native interactions, by avoiding the use of the subroutine *mod_vdw* (file mod_vdw.f) and calculate those directly from the main routine *egb*.

6.3 Future work

In this section a brief description of some recent application of the EVB-SBP method on a protein kinase will be provided. The large-scale conformational transitions of protein kinases represent an issue of great biological and pharmacological relevance. Protein-kinases (PK) transfer a phosphate group from ATP to a specific protein residue, modifying the activity of the target protein. PK are involved in cellular signaling and several vital biochemical functions [348], but deregulated PK have been linked to cancer, diabetes and inflammation, making them attractive targets for drug design. Conformational transition plays a central role in the phosphorylation mechanism. Kinases adopt at least two extreme conformations: An open state that is maximally active and one or more closed states that show minimal activity [349]. Structural insights into the open and the closed states of kinases have been gained from crystal structures of the same protein in different conformations [350, 351]. In several PK, two different ‘closed’ conformations have been found. In the first type, the Asp side chain in the conserved Asp-Phe-Gly (DFG) motif is rotated out of the ATP binding site. In the

second type, the conserved salt bridge that anchors the αC-helix in active structures is broken and the helix moves away from the ATP-binding site. However, a full understanding of the atomistic details of the interplay between the open and the various closed states is still missing.

This work focuses on a specific protein: cellular sarcoma (c-Src) tyrosine kinase. The importance of this kinase is related to its intrinsic pharmaceutical importance, due to its involvement in chronic myelogenous leukemia [352]. Several crystal structures in different conformations are available in the protein data bank: active (PDB 1Y57), intermediate-inactive (PDB 2SRC) and inactive (PDB 2OIQ). The aim is to elucidate the molecular pathways that connect the open (active) conformation with the two possible closed (intermediate, inactive) conformations.

6.3.1 Preliminary results

Structure-based potentials have been used to describe the native basin dynamics of c-Src. The novelty of this application involved the implementation of a different parameterisation procedure, previously described (Section 3.7). In this specific case, the well depth of the native LJ 12-6 function is not taken from the Amber parm99 force field and subsequently rescaled, but directly calculated from all-atom force field simulations. Short (5ns) molecular dynamics simulations were performed for each conformation (active, intermediate and inactive) using the NAMD package [353] and the Amber force field parm99SB [248]. Each structure, consisting of 269 residues, was solvated with approximately ~20,000 TIP3P water and 4 Na+ ions were introduced to reach charge neutrality. Periodic boundary conditions and the particle mesh Ewald method (to account for long range electrostatics) were used. Bonds involving hydrogens were constrained using the SHAKE algorithm [177] and an integration step of 2 fs was used. Simulations were performed in the NVT ensemble at 300 K using the Langevin thermostat.

Using crystallographic reference structures, an initial native interaction list was generated for each conformation with a distance cut-off of 4 Å. This short cut-off was used to reduce the number of native interactions and increase computational efficiency, while reproducing the FF reference data. A further reduction of the native interaction list was achieved by eliminating those atom pairs whose average distance in the FF simulation was beyond 4 Å. As a result, the number of native interactions (N) defined

for each state were as follows: active *N*=2801; intermediate *N*=2716; inactive *N*=2342. Langevin dynamics simulations at 300 K and low friction (γ=5 ps^{-1}) were performed for 5 ns.

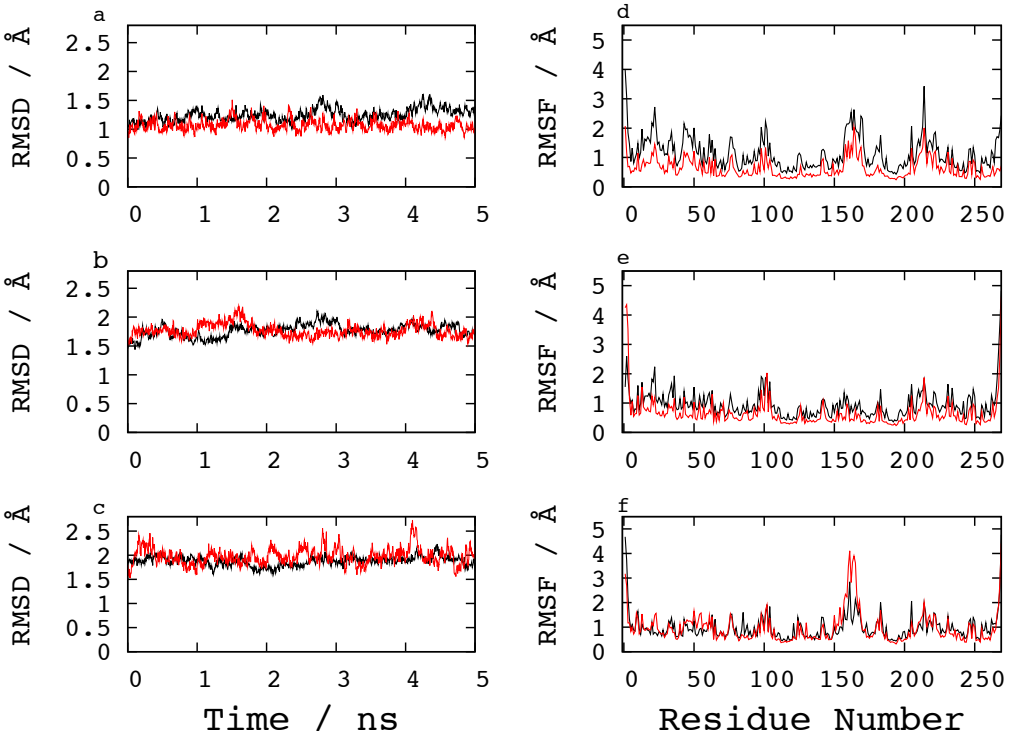

Figure 6.1 Root mean square deviation of atomic positions for c-SRC in different conformations: (a) active (FF=1.24 ± 0.1 Å; SBP=1.07 ± 0.1 Å); (b) intermediate (FF=1.76 ± 0.1 Å; SBP=1.77 ± 0.1 Å), (c) inactive (FF=1.88 ± 0.1 Å; SBP=1.98 ± 0.2 Å) along the trajectory (left) and root mean square fluctuations of residues around the average structure (right) calculated using the structure-based potential (red line) and force field (black line).

The root mean square deviation (RMSD) of the Cartesian coordinates of all atoms in each conformation (active, intermediate, inactive) simulated using SBP is very close to FF results (Fig. 6.1). Positional fluctuations of residues from the respective average structure (RMSF) along the sequence were also reproduced well. The new parameterisation method thus performs very well to reproduce structural and dynamical fluctuations of all-atom force field simulations in the native basins.

Future work will include the parameterisation of the unified potential energy surface by coupling the individual SBP encoding for the active, intermediate-inactive and inactive conformations. On the parameterised energy surface, extensive sampling will be performed by the metadynamic technique [218].

Bibliography

1. Watson, J.D. and F.H.C. Crick, *Molecular Structure of Nucleic Acids - a Structure for Deoxyribose Nucleic Acid.* Nature, 1953. **171**(4356), 737-738.
2. Watson, J.D. and F.H.C. Crick, *The Structure of DNA.* Cold Spring Harbor Symposia on Quantitative Biology, 1953. **18**, 123-131.
3. Adams, J., *The proteasome: structure, function, and role in the cell.* Cancer. Treat. Rev., 2003. 29 Suppl 1, 3-9.
4. Sadre-Bazzaz, K., et al., *Structure of a Blm10 complex reveals common mechanisms for proteasome binding and gate opening.* Molecular Cell, 2010. **37**(5), 728-35.
5. Wu, B., et al., *Structural Insight into the Sequence Dependence of Nucleosome Positioning.* Structure, 2010. **18**(4), 528-536.
6. Konig, H., et al., *Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation.* Cell, 2007. **131**(4), 718-29.
7. Richarz, R., H. Tschesche, and K. Wuthrich, *Structural characterization by nuclear magnetic resonance of a reactive-site 13carbon-labelled basic pancreatic trypsin inhibitor with the peptide bond Arg-39--Ala-40 cleaved and Arg-39 removed.* Eur. J. Biochem., 1979. **102**(2), 563-71.
8. Perutz, M.F., *X-ray analysis of hemoglobin.* Science, 1963. **140**(3569), 863-9.
9. Brunger, A.T. and P.D. Adams, *Molecular dynamics applied to X-ray structure refinement.* Acc. Chem. Res., 2002. **35**(6), 404-12.
10. Linge, J.P., et al., *Refinement of protein structures in explicit solvent.* Proteins: Struct. Funct. Bioinf., 2003. **50**(3), 496-506.
11. Stryer, L., *Biochemistry.* W.H. Freeman and Company, New York, 1995.
12. Lehninger Albert, D.L.N., Michael M Cox, *Principles of Biochemistry.*
13. Breaker, R.R., *DNA aptamers and DNA enzymes.* Curr. Opin. Chem. Biol., 1997. **1**(1), 26-31.
14. Breaker, R.R., *DNA enzymes.* Nat. Biotechnol., 1997. **15**(5), 427-31.
15. Willner, I., et al., *DNAzymes for sensing, nanobiotechnology and logic gate applications.* Chem. Soc. Rev., 2008. **37**(6), 1153-65.
16. Mastroyiannopoulos, N.P., J.B. Uney, and L.A. Phylactou, *The application of ribozymes and DNAzymes in muscle and brain.* Molecules, 2010. **15**(8), 5460-72.
17. Achenbach, J.C., et al., *DNAzymes: From creation in vitro to application in vivo.* Curr. Pharm. Biotechnol., 2004. **5**(4), 321-336.
18. Walter, N.G., et al., *In the fluorescent spotlight: global and local conformational changes of small catalytic RNAs.* Biopolymers, 2001. **61**(3), 224-42.
19. Gesteland, R.F., T.R. Cech, and J.F.E. Atkins, *The RNA World 2nd ed.* Cold Spring Harbor Laboratory Press, 1999.
20. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96), 223-30.
21. Zhang, C.T., *Relations of the numbers of protein sequences, families and folds.* Protein Engineering, 1997. **10**(7), 757-761.
22. Oberai, A., et al., *A limited universe of membrane protein families and folds.* Protein Sci. , 2006. **15**(7), 1723-1734.
23. Dunker, A.K. and V.N. Uversky, *Signal transduction via unstructured protein conduits.* Nat. Chem. Biol., 2008. **4**(4), 229-30.
24. Sigalov, A.B., *Protein intrinsic disorder and oligomericity in cell signaling.* Mol. Biosyst., 2010. **6**(3), 451-61.
25. Mittag, T., L.E. Kay, and J.D. Forman-Kay, *Protein dynamics and conformational disorder in molecular recognition.* J. Mol. Recognit., 2010. **23**(2), 105-16.
26. Wright, P.E. and H.J. Dyson, *Linking folding and binding.* Curr. Opin. Struct. Biol., 2009. **19**(1), 31-8.
27. Sudarsan, N., et al., *An mRNA structure in bacteria that controls gene expression by binding lysine.* Genes & Development, 2003. **17**(21), 2688-2697.
28. Winkler, W.C., et al., *An mRNA structure that controls gene expression by binding S-adenosylmethionine.* Nat. Struct. Biol., 2003. **10**(9), 701-707.
29. Wand, A.J., *Dynamic activation of protein function: A view emerging from NMR spectroscopy.* Nat. Struct. Biol., 2001. **8**(11), 926-931.

30. Boehr, D.D., H.J. Dyson, and P.E. Wright, *An NMR perspective on enzyme dynamics.* Chem. Rev., 2006. **106**(8), 3055-3079.
31. Dyson, H.J. and P.E. Wright, *Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance.* Unfolded Proteins, 2002. **62**, 311-340.
32. Eliezer, D., et al., *Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding.* Nat. Struct. Biol., 1998. **5**(2), 148-155.
33. Wright, P.E., *What can two-dimensional NMR tell us about proteins.* Trends in Biochemical Sciences, 1989. **14**(7), 255-260.
34. Frank, A.T., et al., *Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition.* Nucleic Acids Res, 2009. **37**(11), 3670-3679.
35. Al-Hashimi, H.M. and N.G. Walter, *RNA dynamics: it is about time.* Curr. Opin. Struct. Biol., 2008. **18**(3), 321-329.
36. Getz, M., et al., *NMR studies of RNA dynamics and structural plasticity using NMR residual dipolar couplings.* Biopolymers, 2007. **86**(5-6), 384-402.
37. Linderstrøm-Lang, K.U, *Proteins and Enzymes.* Stanford University Press, 1952. **6**(Lane Medical Lectures, Stanford University Publications, University Series, Medical Sciences).
38. Branden, C.I., Tooze J., *Introduction to Protein Structure.* Garland Publishing Inc, 1998.
39. Pauling, L. and R.B. Corey, *Configuration of Polypeptide Chains.* Nature, 1951. **168**(4274), 550-551.
40. Pauling, L., R.B. Corey, and H.R. Branson, *The Structure of Proteins - 2 Hydrogen-Bonded Helical Configurations of the Polypeptide Chain.* Proc. Nat. Acad. Sci. USA, 1951. **37**(4), 205-211.
41. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J. Mol. Biol., 1995. **247**(4), 536-40.
42. Orengo, C.A., et al., *The CATH Database provides insights into protein structure/function relationships.* Nucleic Acids Res, 1999. **27**(1), 275-9.
43. Levitt, M. and C. Chothia, *Structural Patterns in Globular Proteins.* Nature, 1976. **261**(5561), 552-558.
44. Nagaswamy, U., et al., *Database of non-canonical base pairs found in known RNA structures.* Nucleic Acids Res, 2000. **28**(1), 375-376.
45. Leontis, N.B., J. Stombaugh, and E. Westhof, *The non-Watson-Crick base pairs and their associated isostericity matrices.* Nucleic Acids Res, 2002. **30**(16), 3497-531.
46. Chandrasekaran, R. and S. Arnott, *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology.* Springer-Verlag, Berlin, 1989.
47. Drew, H.R., et al., *Structure of a B-DNA dodecamer: conformation and dynamics.* Proc. Natl. Acad. Sci. USA, 1981. **78**(4), 2179-83.
48. Wang, A.H., et al., *Molecular structure of a left-handed double helical DNA fragment at atomic resolution.* Nature, 1979. **282**(5740), 680-6.
49. Horvath, M.P. and S.C. Schultz, *DNA G-quartets in a 1.86 A resolution structure of an Oxytricha nova telomeric protein-DNA complex.* J. Mol. Biol., 2001. **310**(2), 367-77.
50. Haider, S., G.N. Parkinson, and S. Neidle, *Crystal structure of the potassium form of an Oxytricha nova G-quadruplex.* J. Mol. Biol., 2002. **320**(2), 189-200.
51. Huppert, J.L. and S. Balasubramanian, *Prevalence of quadruplexes in the human genome.* Nucleic Acids Res, 2005. **33**(9), 2908-16.
52. Burge, S., et al., *Quadruplex DNA: sequence, topology and structure.* Nucleic Acids Res, 2006. **34**(19), 5402-15.
53. Neidle, S. and G. Parkinson, *Telomere maintenance as a target for anticancer drug discovery.* Nat. Rev. Drug. Discov., 2002. **1**(5), 383-93.
54. Lipps, H.J. and D. Rhodes, *G-quadruplex structures: in vivo evidence and function.* Trends. Cell Biol., 2009. **19**(8), 414-22.
55. Gueron, M. and J.L. Leroy, *The i-motif in nucleic acids.* Curr. Opin Struct. Biol., 2000. **10**(3), 326-331.
56. Holliday, R., *Molecular aspects of genetic exchange and gene conversion.* Genetics, 1974. **78**(1), 273-87.
57. Stahl, F.W., *The Holliday junction on its thirtieth anniversary.* Genetics, 1994. **138**(2), 241-6.
58. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J. Mol. Biol. , 1977. **112**(3), 535-42.
59. Berman, H.M., et al., *The Nucleic-Acid Database - a Comprehensive Relational Database of 3-Dimensional Structures of Nucleic-Acids.* Biophys. J., 1992. **63**(3), 751-759.

60. Moore, P.B., *Structural motifs in RNA.* Annu. Rev. Biochem., 1999. **68**(1), 287-300.
61. Doty, P., et al., *Secondary Structure in Ribonucleic Acids.* Proc. Natl. Acad. Sci. USA, 1959. **45**(4), 482-99.
62. Leontis, N.B., A. Lescoute, and E. Westhof, *The building blocks and motifs of RNA architecture.* Curr. Opin. Struct. Biol., 2006. **16**(3), 279-87.
63. Karplus, M. and J.A. McCammon, *Dynamics of Proteins.* Scientific American, 1986. **254**(4), 42-51.
64. McCammon, J.A. and S.C. Harvey, *Dynamics of Proteins and Nucleic Acids.* Cambridge University Press, 1987.
65. Gerstein, M. and C. Chothia, *Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase.* J. Mol. Biol., 1991. **220**(1), 133-49.
66. Cheng, Y.H., Y.K. Zhang, and J.A. McCammon, *How does activation loop phosphorylation modulate catalytic activity in the cAMP-dependent protein kinase: A theoretical study.* Protein Sci. , 2006. **15**(4), 672-683.
67. Frauenfelder, H., S.G. Sligar, and P.G. Wolynes, *The energy landscapes and motions of proteins.* Science, 1991. **254**(5038), 1598-603.
68. Bahar, I., et al., *Global dynamics of proteins: bridging between structure and function.* Annu. Rev. Biophys., 2010. **39**, 23-42.
69. Gerstein, M. and W. Krebs, *A database of macromolecular motions.* Nucleic Acids Res, 1998. **26**(18), 4280-90.
70. Benitah, J.P., et al., *Molecular motions within the pore of voltage-dependent sodium channels revealed by engineered disulfide trapping.* Biophys. J., 1997. **72**(2), 603-613.
71. Pisliakov, A.V., et al., *Enzyme millisecond conformational dynamics do not catalyze the chemical step.* Proc. Natl. Acad. Sci. USA, 2009. **106**(41), 17359-17364.
72. Howard, J., *Mechanical Signaling in Networks of Motor and Cytoskeletal Proteins.* Annu. Rev. Biophys., 2009. **38**, 217-234.
73. Guerin, T., et al., *Coordination and collective properties of molecular motors: theory.* Curr. Opin. Cell Biol., 2010. **22**(1), 14-20.
74. Kern, D., E.Z. Eisenmesser, and M. Wolf-Watz, *Enzyme dynamics during catalysis measured by NMR spectroscopy.* Methods. Enzymol., 2005. **394**, 507-24.
75. Hammes-Schiffer, S. and S.J. Benkovic, *Relating protein motion to catalysis.* Annu. Rev. Biochem., 2006. **75**, 519-41.
76. Henzler-Wildman, K.A., et al., *A hierarchy of timescales in protein dynamics is linked to enzyme catalysis.* Nature, 2007. **450**(7171), 913-916.
77. Kamerlin, S.C. and A. Warshel, *At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?* Proteins: Struct. Funct. Bioinf., 2010. **78**(6), 1339-75.
78. Karplus, M., *Role of conformation transitions in adenylate kinase.* Proc. Natl. Acad. Sci. USA, 2010. **107**(17), E71; author reply E72.
79. Boehr, D.D., R. Nussinov, and P.E. Wright, *The role of dynamic conformational ensembles in biomolecular recognition.* Nat. Chem. Biol., 2009. **5**(11), 789-796.
80. Kalodimos, C.G., R. Boelens, and R. Kaptein, *Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the Lac repressor system.* Chem. Rev., 2004. **104**(8), 3567-3586.
81. Prevost, C., M. Takahashi, and R. Lavery, *Deforming DNA: from physics to biology.* ChemPhysChem, 2009. **10**(9-10), 1399-404.
82. Gueron, M. and J.L. Leroy, *Studies of base pair kinetics by NMR measurement of proton exchange.* Methods Enzymol, 1995. **261**, 383-413.
83. Varnai, P. and R. Lavery, *Base flipping in DNA: Pathways and energetics studied with molecular dynamic simulations.* J. Am. Chem. Soc., 2002. **124**(25), 7272-7273.
84. Cubero, E., et al., *Observation of spontaneous base pair breathing events in the molecular dynamics simulation of a difluorotoluene-containing DNA oligonucleotide.* J. Am. Chem. Soc., 1999. **121**(37), 8653-8654.
85. Banavali, N.K. and A.D. MacKerell, *Free energy and structural pathways of base flipping in a DNA GCGC containing sequence.* J. Mol. Biol., 2002. **319**(1), 141-160.
86. Strauss, J.K. and L.J. Maher, 3rd, *DNA bending by asymmetric phosphate neutralization.* Science, 1994. **266**(5192), 1829-34.
87. Cluzel, P., et al., *DNA: an extensible molecule.* Science, 1996. **271**(5250), 792-4.
88. Geiselmann, J., *The role of DNA conformation in transcriptional initiation and activation in Escherichia coli.* Biol. Chem., 1997. **378**(7), 599-607.
89. Williamson, J.R., *Induced fit in RNA-protein recognition.* Nat. Struct. Biol., 2000. **7**(10), 834-7.

90. Hall, K.B., *RNA in motion.* Curr. Opin. Chem. Biol., 2008. **12**(6), 612-8.
91. Franklin, R.E. and R.G. Gosling, *The Structure of Sodium Thymonucleate Fibres .2. The Cylindrically Symmetrical Patterson Function.* Acta Crystallographica, 1953. **6**(8-9), 678-685.
92. Muirhead, H. and M.F. Perutz, *Structure of Haemoglobin. A Three-Dimensional Fourier Synthesis of Reduced Human Haemoglobin at 5-5 a Resolution.* Nature, 1963. **199**, 633-8.
93. Kendrew, J.C., *Architecture of a protein molecule.* Nature, 1958. **182**(4638), 764-7.
94. Holmgren, A., et al., *Three-dimensional structure of Escherichia coli thioredoxin-S2 to 2.8 A resolution.* Proc. Natl. Acad. Sci. USA, 1975. **72**(6), 2305-9.
95. Doniach, S., *Changes in biomolecular conformation seen by small angle X-ray scattering.* Chem. Rev., 2001. **101**(6), 1763-78.
96. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, 1958. **181**(4610), 662-6.
97. Simonovic, M. and T.A. Steitz, *Peptidyl-CCA deacylation on the ribosome promoted by induced fit and the O3'-hydroxyl group of A76 of the unacylated A-site tRNA.* RNA, 2008. **14**(11), 2372-8.
98. Groll, M., et al., *Crystal structure of the 20 S proteasome:TMC-95A complex: a non-covalent proteasome inhibitor.* J. Mol. Biol. , 2001. **311**(3), 543-8.
99. Lerch, T.F., Q. Xie, and M.S. Chapman, *The structure of adeno-associated virus serotype 3B (AAV-3B): Insights into receptor binding and immune evasion.* Virology, 2010. **403**(1), 26-36.
100. Weiss, S., *Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy.* Nat. Struct. Biol., 2000. **7**(9), 724-9.
101. Haustein, E. and P. Schwille, *Single-molecule spectroscopic methods.* Curr. Opin. Struct. Biol., 2004. **14**(5), 531-40.
102. Ha, T., *Single-molecule fluorescence resonance energy transfer.* Methods, 2001. **25**(1), 78-86.
103. Joo, C., et al., *Advances in single-molecule fluorescence methods for molecular biology.* Annu. Rev. Biochem., 2008. **77**, 51-76.
104. Truong, K. and M. Ikura, *The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo.* Curr. Opin. Struct. Biol., 2001. **11**(5), 573-578.
105. Sako, Y., S. Minoghchi, and T. Yanagida, *Single-molecule imaging of EGFR signalling on the surface of living cells.* Nat. Cell Biol., 2000. **2**(3), 168-172.
106. Chung, E.W., et al., *Hydrogen exchange properties of proteins in native and denatured states monitored by mass spectrometry and NMR.* Prot. Sci, 1997. **6**(6), 1316-24.
107. Busenlehner, L.S. and R.N. Armstrong, *Insights into enzyme structure and dynamics elucidated by amide H/D exchange mass spectrometry.* Arch. Biochem. Biophys., 2005. **433**(1), 34-46.
108. Wales, T.E. and J.R. Engen, *Hydrogen exchange mass spectrometry for the analysis of protein dynamics.* Mass. Spectrom. Rev., 2006. **25**(1), 158-70.
109. Schunemann, V. and H. Winkler, *Structure and dynamics of biomolecules studied by Mossbauer spectroscopy.* Reports on Progress in Physics, 2000. **63**(3), 263-353.
110. Thomas, G.J., Jr., *Raman spectroscopy of protein and nucleic acid assemblies.* Annu. Rev. Biophys. Biomol. Struct., 1999. **28**, 1-27.
111. Zanni, M.T. and R.M. Hochstrasser, *Two-dimensional infrared spectroscopy: a promising new method for the time resolution of structures.* Curr. Opin. Struct. Biol., 2001. **11**(5), 516-22.
112. Parak, F.G., *Proteins in action: the physics of structural fluctuations and conformational changes.* Curr. Opin. Struct. Biol., 2003. **13**(5), 552-7.
113. Sage, J.T., et al., *Long-range reactive dynamics in myoglobin.* Phys. Rev. Lett., 2001. **86**(21), 4966-4969.
114. Benevides, J.M., S.A. Overman, and G.J. Thomas, *Raman, polarized Raman and ultraviolet resonance Raman spectroscopy of nucleic acids and their complexes.* Journal of Raman Spectroscopy, 2005. **36**(4), 279-299.
115. Tuma, R., *Raman spectroscopy of proteins: from peptides to large assemblies.* Journal of Raman Spectroscopy, 2005. **36**(4), 307-319.
116. Hunt, N.T., *2D-IR spectroscopy: ultrafast insights into biomolecule structure and function.* Chem. Soc. Rev., 2009. **38**(7), 1837-1848.
117. Bragg, W.L., *The analysis of crystals by the X-ray spectrometer.* Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character, 1914. **89**(613), 468-489.
118. Rouse, K.D., B.T.M. Willis, and A.W. Pryor, *Anharmonic Contributions to Debye-Waller Factors of Uo2.* Acta Crystallographica Section B-Structural Crystallography and Crystal Chemistry, 1968. **B 24**, 117-122.

119. Waller, I. and R.W. James, *On the temperature factors of X-ray reflexion for sodium and chlorine in the rock-salt crystal.* Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character, 1927. **117**(776), 214-223.
120. Aue, W.P., E. Bartholdi, and R.R. Ernst, *2-Dimensional Spectroscopy - Application to Nuclear Magnetic-Resonance.* J. Chem. Phys., 1976. **64**(5), 2229-2246.
121. Overhauser, A.W., *Polarization of Nuclei in Metals.* Phys. Rev., 1953. **92**(2), 411-415.
122. Zuiderweg, E.R., R. Kaptein, and K. Wuthrich, *Sequence-specific resonance assignments in the 1H nuclear-magnetic-resonance spectrum of the lac repressor DNA-binding domain 1-51 from Escherichia coli by two-dimensional spectroscopy.* Eur. J. Biochem., 1983. **137**(1-2), 279-92.
123. Kay, L.E., *Protein dynamics from NMR.* Nat. Struct. Biol., 1998. **5 Suppl**, 513-7.
124. Al-Hashimi, H.M., *Beyond static structures of RNA by NMR: Folding, refolding, and dynamics at atomic resolution.* Biopolymers, 2007. **86**(5-6), 345-347.
125. Kay, L.E., D.A. Torchia, and A. Bax, *Backbone Dynamics of Proteins as Studied by N-15 Inverse Detected Heteronuclear Nmr-Spectroscopy - Application to Staphylococcal Nuclease.* Biochemistry, 1989. **28**(23), 8972-8979.
126. Lipari, G. and A. Szabo, *Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .1. Theory and Range of Validity.* J. Am. Chem. Soc., 1982. **104**(17), 4546-4559.
127. Tolman, J.R., et al., *NMR evidence for slow collective motions in cyanometmyoglobin.* Nat. Struct. Biol., 1997. **4**(4), 292-297.
128. Robinson, C.V., et al., *Probing the nature of noncovalent interactions by mass spectrometry. A study of protein-CoA ligand binding and assembly.* J. Am. Chem. Soc., 1996. **118**(36), 8646-8653.
129. Garcia, R.A., D. Pantazatos, and F.J. Villarreal, *Hydrogen/deuterium exchange mass spectrometry for investigating protein-ligand interactions.* Assay and Drug Development Technologies, 2004. **2**(1), 81-91.
130. Fiaux, J., et al., *NMR analysis of a 900K GroEL-GroES complex.* Nature, 2002. **418**(6894), 207-211.
131. Fernandez, C. and G. Wider, *TROSY in NMR studies of the structure and function of large biological macromolecules.* Curr. Opin. Struct. Biol., 2003. **13**(5), 570-580.
132. Cabrita, L.D., et al., *Probing ribosome-nascent chain complexes produced in vivo by NMR spectroscopy.* Proc. Natl. Acad. Sci. USA, 2009. **106**(52), 22239-44.
133. Feynman, R.P., R.B. Leighton, and M. Sands, *The Feynman Lectures in Physics* Addison-Wesley, 1963. **1**(3-6).
134. Allen, M.P. and D.J. Tildesley, *Computer Simulation of liquids.* Oxford: University Press, 1989.
135. Leach, A.R., *Molecular Modeling: Principles and Applications.* Prenitice Hall, 2001.
136. Alder, B.J. and T.E. Wainwright, *Phase Transition for a Hard Sphere System.* J. Chem. Phys., 1957. **27**(5), 1208-1209.
137. Alder, B.J. and T.E. Wainwright, *Molecular Motions.* Scientific American, 1959. **201**(4), 113-126.
138. Alder, B.J. and T.E. Wainwright, *Studies in Molecular Dynamics .1. General Method.* J. Chem. Phys., 1959. **31**(2), 459-466.
139. McCammon, J.A., B.R. Gelin, and M. Karplus, *Dynamics of Folded Proteins.* Nature, 1977. **267**(5612), 585-590.
140. Levitt, M., *Computer simulation of DNA double-helix dynamics.* Cold Spring Harb Symp Quant Biol, 1983. **47 Pt 1**, 251-62.
141. Hehre, W.J., et al., *Ab initio molecular orbital theory.* John wiley New York, 1986.
142. Szabo, A., Ostlund, N. S., *Modern Quantum Chemistry.* New York : McGraw Hill, 1989.
143. Barlett, R.J., Stanton, J. F. , *Applications of post-Hartree-Fock methods: A tutorial.In : Reviews in Computational Chemistry.* VCH Publishers, Inc New York, 1994. **5**, 65-169.
144. Stewart, J.J.P., *Semi-empirical molecular orbitals methods. In : Reviews in Computational Chemistry.* VCH Publishers, Inc New York, 1990. **1**, 209-220.
145. Zerner, R.W., *Semi-empirical molecular orbitals methods. In : Reviews in Computational Chemistry.* VCH Publishers, Inc New York, 1991. **2**, 313-365.
146. Dewar, M.J.S. and W. Thiel, *Ground-States of Molecules .38. Mndo Method - Approximations and Parameters.* J. Am. Chem. Soc., 1977. **99**(15), 4899-4907.
147. Dewar, M.J.S., et al., *The Development and Use of Quantum-Mechanical Molecular-Models .76. Am1 - a New General-Purpose Quantum-Mechanical Molecular-Model.* J. Am. Chem. Soc., 1985. **107**(13), 3902-3909.

148. Stewart, J.J.P., *Optimization of Parameters for Semiempirical Methods .1. Method.* J. Comput. Chem., 1989. **10**(2), 209-220.
149. Stewart, J.J.P., *Comparison of the accuracy of semiempirical and some DFT methods for predicting heats of formation.* J. Mol. Model., 2004. **10**(1), 6-12.
150. Brooks, I., C. L. , M. Karplus, and B.M. Pettitt, *Proteins: a theoretical perspective of dynamics, structure and thermodynamics.* Wiley-Interscience, 1988.
151. McCammon, J.A. and S.C. Harvey, *Dynamics of protein and nucleic acids.* Cambridge University Press, 1987.
152. Perez, A., et al., *Refinenement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers.* Biophys. J., 2007. **92**(11), 3817-3829.
153. Wang, J.M., P. Cieplak, and P.A. Kollman, *How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?* J. Comput. Chem., 2000. **21**(12), 1049-1074.
154. Brooks, B.R., et al., *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* J. Comput. Chem., 1983. **4**(2), 187-217.
155. Jorgensen, W.L. and J. Tirado-Rives, *The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin.* J. Am. Chem. Soc., 1988. **110**(6), 1657-1666.
156. Ponder, J.W., et al., *Current status of the AMOEBA polarizable force field.* J. Phys. Chem. B, 2010. **114**(8), 2549-64.
157. Ren, P. and J.W. Ponder, *Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations.* J. Comput. Chem., 2002. **23**(16), 1497-506.
158. Duan, Y., et al., *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations.* J. Comput. Chem., 2003. **24**(16), 1999-2012.
159. Cheatham, T.E., et al., *Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins.* J. Am. Chem. Soc., 1995. **117**(14), 4193-4194.
160. Darden, T.A., D. York, and L. Pedersen, *Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems.* J. Chem. Phys, 1993. **98**, 10089-10092.
161. Banas P., et al., *Performance of Molecular Mechanics Force Fields for RNA Simulations: Stability of UUCG and GNRA Hairpins.* Journal of Chemical Theory and Computation, 2010. **6**(12), 3836-3849.
162. Berendsen, H.J.C., J.R. Grigera, and T.P. Straatsma, *The Missing Term in Effective Pair Potentials.* J. Phys. Chem., 1987. **91**(24), 6269-6271.
163. Jorgensen, W.L., et al., *Comparison of Simple Potential Functions for Simulating Liquid Water.* J. Chem. Phys., 1983. **79**(2), 926-935.
164. Mahoney, M. and W.L. Jorgensen, *A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions.* J. Chem. Phys., 2000. **112**(20), 8910-8922.
165. Mark, P. and L. Nilsson, *Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K.* J. Phys. Chem. A, 2001. **105**(43), 9954-9960.
166. Marchi, M., F. Sterpone, and M. Ceccarelli, *Water rotational relaxation and diffusion in hydrated lysozyme.* J. Am. Chem. Soc. , 2002. **124**(23), 6787-91.
167. Ferguson, D.M., *Parameterization and Evaluation of a Flexible Water Model.* J. Comput. Chem., 1995. **16**(4), 501-511.
168. Sprik, M. and M.L. Klein, *A Polarizable Model for Water Using Distributed Charge Sites.* J. Chem. Phys., 1988. **89**(12), 7556-7560.
169. Roux, B. and T. Simonson, *Implicit solvent models.* Biophysical Chemistry, 1999. **78**(1-2), 1-20.
170. Still, W.C., et al., *Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics.* J. Am. Chem. Soc., 1990. **112**(16), 6127-6129.
171. Born, M., *Volumen and hydratationswarme der ionen.* Zeitschrift Fur Physikalische Chemie-International Journal of Research in Physical Chemistry & Chemical Physics, 1920. **1**, 45-48.
172. Tsui, V. and D.A. Case, *Theory and applications of the generalized Born solvation model in macromolecular simulations.* Biopolymers, 2000. **56**(4), 275-91.
173. Fogolari, F., A. Brigo, and H. Molinari, *The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology.* J. Mol. Recognit., 2002. **15**(6), 377-92.
174. R.G.Palmer, *Broken ergodicity.* Adv. Phys., 1982. **31**(6), 669-735.
175. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules.* Nat. Struct. Biol., 2002. **9**(9), 646-652.

176. Vangunsteren, W.F. and H.J.C. Berendsen, *Algorithms for Macromolecular Dynamics and Constraint Dynamics.* Molecular Physics, 1977. **34**(5), 1311-1327.
177. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes.* J. Comput. Phys., 1977. **23**(3), 327-341.
178. *Hess, B., et al., LINCS: A linear constraint solver for molecular simulations.* J. Comput. Phys. *1997. 18(12): p. 1463-1472*
179. Voelz, V.A., et al., *Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39).* J. Am. Chem. Soc., 2010. **132**(5), 1526-8.
180. Brooks, C.L., et al., *Chemical physics of protein folding.* Proc. Natl. Acad. Sci. USA, 1998. **95**(19), 11037-11038.
181. Bryngelson, J.D. and P.G. Wolynes, *Intermediates and Barrier Crossing in a Random Energy-Model (with Applications to Protein Folding).* J. Phys. Chem., 1989. **93**(19), 6902-6915.
182. Onuchic, J.N., Z. Luthey-Schulten, and P.G. Wolynes, *Theory of protein folding: The energy landscape perspective.* Annu. Rev. Phys. Chem., 1997. **48**, 545-600.
183. Hansmann, U.H.E. and Y. Okamoto, *Numerical comparisons of three recently proposed algorithms in the protein folding problem.* J. Comput. Chem., 1997. **18**(7), 920-933.
184. Tejero, R., et al., *Simulated annealing with restrained molecular dynamics using CONGEN: energy refinement of the NMR solution structures of epidermal and type-alpha transforming growth factors.* Protein Sci. , 1996. **5**(4), 578-92.
185. Elber, R. and M. Karplus, *Enhanced Sampling in Molecular-Dynamics - Use of the Time-Dependent Hartree Approximation for a Simulation of Carbon-Monoxide Diffusion through Myoglobin.* J. Am. Chem. Soc., 1990. **112**(25), 9161-9175.
186. Keasar, C. and R. Elber, *Homology as a Tool in Optimization Problems - Structure Determination of 2d Heteropolymers.* J. Phys. Chem., 1995. **99**(29), 11550-11556.
187. Simmerling, C., J.L. Miller, and P.A. Kollman, *Combined locally enhanced sampling and Particle Mesh Ewald as a strategy to locate the experimental structure of a nonhelical nucleic acid.* J. Am. Chem. Soc., 1998. **120**(29), 7149-7155.
188. Joseph-McCarthy, D., et al., *Use of MCSS to design small targeted libraries: application to picornavirus ligands.* J. Am. Chem. Soc., 2001. **123**(51), 12758-69.
189. Swendsen, R.H. and J.S. Wang, *Replica Monte Carlo simulation of spin glasses.* Phys. Rev. Lett., 1986. **57**(21), 2607-2609.
190. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chem. Phys. Lett., 1999. **314**(1-2), 141-151.
191. Sanbonmatsu, K.Y. and A.E. Garcia, *Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics.* Proteins: Struct. Funct. Gen., 2002. **46**(2), 225-234.
192. Pitera, J.W. and W. Swope, *Understanding folding and design: Replica-exchange simulations of "Trp-cage" fly miniproteins.* Proc. Natl. Acad. Sci. USA, 2003. **100**(13), 7587-7592.
193. Faradjian, A.K. and R. Elber, *Computing time scales from reaction coordinates by milestoning.* J. Chem. Phys., 2004. **120**(23), 10880-10889.
194. Dellago, C., et al., *Transition path sampling and the calculation of rate constants.* J. Chem. Phys., 1998. **108**(5), 1964-1977.
195. Dellago, C., P.G. Bolhuis, and D. Chandler, *Efficient transition path sampling: Application to Lennard-Jones cluster rearrangements.* J. Chem. Phys., 1998. **108**(22), 9236-9245.
196. Taketomi, H., Y. Ueda, and N. Go, *Studies on Protein Folding, Unfolding and Fluctuations by Computer-Simulation .1. Effect of Specific Amino-Acid Sequence Represented by Specific Inter-Unit Interactions.* International Journal of Peptide and Protein Research, 1975. **7**(6), 445-459.
197. Karanicolas, J. and C.L. Brooks, 3rd, *The origins of asymmetry in the folding transition states of protein L and protein G.* Protein Sci. , 2002. **11**(10), 2351-61.
198. Bahar, I. and R.L. Jernigan, *Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation.* J. Mol. Biol. , 1997. **266**(1), 195-214.
199. Saunders, J.A. and H.A. Scheraga, *Ab initio structure prediction of two alpha-helical oligomers with a multiple-chain united-residue force field and global search.* Biopolymers, 2003. **68**(3), 300-17.
200. Hyeon, C. and D. Thirumalai, *Mechanical unfolding of RNA hairpins.* Proc. Natl. Acad. Sci. USA, 2005. **102**(19), 6789-94.
201. Sorin, E.J., et al., *Does native state topology determine the RNA folding mechanism?* J. Mol. Biol. , 2004. **337**(4), 789-97.

202. Trovato, F. and V. Tozzini, *Supercoiling and local denaturation of plasmids with a minimalist DNA model.* J. Phys. Chem. B, 2008. **112**(42), 13197-200.
203. Voltz, K., et al., *Coarse-grained force field for the nucleosome from self-consistent multiscaling.* J. Comput. Chem., 2008. **29**(9), 1429-39.
204. Best, R.B., Y.G. Chen, and G. Hummer, *Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor.* Structure, 2005. **13**(12), 1755-1763.
205. Okazaki, K., et al., *Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations.* Proc. Natl. Acad. Sci. USA, 2006. **103**(32), 11844-11849.
206. Kirkwood, J.G., *Statistical mechanics of liquid solutions.* Chem. Rev., 1936. **19**(3), 275-307.
207. Zwanzig, R.W., *High-Temperature Equation of State by a Perturbation Method .1. Nonpolar Gases.* J. Chem. Phys., 1954. **22**(8), 1420-1426.
208. Seeliger, D. and B.L. de Groot, *Protein thermostability calculations using alchemical free energy simulations.* Biophys. J., 2010. **98**(10), 2309-16.
209. Bash, P.A., et al., *Free-Energy Calculations by Computer-Simulation.* Science, 1987. **236**(4801), 564-568.
210. Torrie, G.M. and J.P. Valleau, *Non-Physical Sampling Distributions in Monte-Carlo Free-Energy Estimation - Umbrella Sampling.* J. Comput. Phys., 1977. **23**(2), 187-199.
211. Kumar, S., et al., *The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules .1. The Method.* J. Comput. Chem., 1992. **13**(8), 1011-1021.
212. Ferrenberg, A.M. and R.H. Swendsen, *Optimized Monte-Carlo Data-Analysis.* Phys. Rev. Lett., 1989. **63**(12), 1195-1198.
213. Kastner, J. and W. Thiel, *Analysis of the statistical error in umbrella sampling simulations by umbrella integration.* J. Chem. Phys., 2006. **124**(23), 234106.
214. Kobrak, M.N., *Systematic and statistical error in histogram-based free energy calculations.* J. Comput. Chem., 2003. **24**(12), 1437-46.
215. Roux, B., *The calculation of the potential of mean force using computer simulations.* Comput. Phys. Commun., 1994. **91**(1-3), 275-282.
216. Best, R.B. and G. Hummer, *Reaction coordinates and rates from transition paths.* Proc. Natl. Acad. Sci. USA, 2005. **102**(19), 6732-6737.
217. Mezei, M., *Adaptive Umbrella Sampling - Self-Consistent Determination of the Non-Boltzmann Bias.* J. Comput. Phys., 1987. **68**(1), 237-248.
218. Laio, A. and M. Parrinello, *Escaping free-energy minima.* Proc. Natl. Acad. Sci. USA, 2002. **99**(20), 12562-12566.
219. Raiteri, P., et al., *Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics.* J. Phys. Chem. B, 2006. **110**(8), 3533-3539.
220. Bussi, G., et al., *Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics.* J. Am. Chem. Soc., 2006. **128**(41), 13435-13441.
221. Warshel, A. and R.M. Weiss, *An Empirical Valence Bond Approach for Comparing Reactions in Solutions and in Enzymes.* J. Am. Chem. Soc., 1980. **102**(20), 6218-6226.
222. Warshel, A. and M. Levitt, *Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme.* J. Mol. Biol. , 1976. **103**(2), 227-49.
223. Chang Y.T, M.W.H., *An Empirical Valence Bond Model for Constructing Global Potential Energy Surfaces for Chemical Reactions of Polyatomic Molecular Systems.* J. Phys. Chem., 1990. **94**, 5884.
224. Tozzini, V., *Coarse-grained models for proteins.* Curr. Opin. Struct. Biol., 2005. **15**(2), 144-150.
225. Knotts, T.A., et al., *A coarse grain model for DNA.* J. Chem. Phys., 2007. **126**(8), 084901.
226. Whitford, P.C., et al., *An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields.* Proteins: Struct. Funct. Bioinf., 2009. **75**(2), 430-441.
227. Whitford, P.C., et al., *Nonlocal Helix Formation Is Key to Understanding S-Adenosylmethionine-1 Riboswitch Function.* Biophys. J., 2009. **96**(2), L7-L9.
228. Bryngelson, J.D., et al., *Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis.* Proteins: Struct. Funct. Gen., 1995. **21**(3), 167-195.
229. Plaxco, K.W., K.T. Simons, and D. Baker, *Contact order, transition state placement and the refolding rates of single domain proteins.* J. Mol. Biol., 1998. **277**(4), 985-994.
230. Leopold, P.E., M. Montal, and J.N. Onuchic, *Protein Folding Funnels - a Kinetic Approach to the Sequence Structure Relationship.* Proc. Natl. Acad. Sci. USA, 1992. **89**(18), 8721-8725.

231. Onuchic, J.N. and P.G. Wolynes, *Theory of protein folding.* Curr. Opin. Struct. Biol., 2004. **14**(1), 70-75.
232. Clementi, C., P.A. Jennings, and J.N. Onuchic, *Prediction of folding mechanism for circular-permuted proteins.* J. Mol. Biol., 2001. **311**(4), 879-890.
233. Shoemaker, B.A., J. Wang, and P.G. Wolynes, *Structural correlations in protein folding funnels.* Proc. Natl. Acad. Sci. USA, 1997. **94**(3), 777-782.
234. Sosnick, T.R. and T. Pan, *Reduced contact order and RNA folding rates.* J. Mol. Biol., 2004. **342**(5), 1359-1365.
235. Zhang, B.W., D. Jasnow, and D.M. Zuckerman, *Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin.* Proc. Natl. Acad. Sci. USA, 2007. **104**(46), 18043-18048.
236. Maragakis, P. and M. Karplus, *Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase.* J. Mol. Biol., 2005. **352**(4), 807-822.
237. Chu, J.W. and G.A. Voth, *Coarse-grained free energy functions for studying protein conformational changes: A double-well network model.* Biophys. J., 2007. **93**(11), 3860-3871.
238. De Marco, G. and P. Varnai, *Molecular simulation of conformational transitions in biomolecules using a combination of structure-based potential and empirical valence bond theory.* Physical Chemistry Chemical Physics, 2009. **11**(45), 10694-10700.
239. Brooks, C.L., 3rd, J.N. Onuchic, and D.J. Wales, *Statistical thermodynamics. Taking a walk on a landscape.* Science, 2001. **293**(5530), 612-3.
240. Ptitsyn, O.B., *Structures of folding intermediates.* Curr. Opin. Struct. Biol., 1995. **5**(1), 74-8.
241. Fersht, A.R., *Characterizing transition states in protein folding: an essential step in the puzzle.* Curr. Opin. Struct. Biol., 1995. **5**(1), 79-84.
242. Bryngelson, J.D. and P.G. Wolynes, *Spin glasses and the statistical mechanics of protein folding.* Proc. Natl. Acad. Sci. USA, 1987. **84**(21), 7524-8.
243. D.A. Case, T.A.D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, and R.L. R.E. Duke, M. Crowley, Ross C. Walker,W. Zhang, K.M. Merz, B.Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F.Wong, F. Paesani, J. Vanicek, X.Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, and P.A. Kollman, University of California, San Francisco, 2008.
244. Morse, P.M., *Diatomic molecules according to the wave mechanics. II. Vibrational levels.* Phys. Rev., 1929. **34**(1), 57-64.
245. Wang, L. and J. Hermans, *Change of bond length in free-energy simulations: Algorithmic improvements, but when is it necessary?* J. Chem. Phys., 1993. **100**(12), 9129-9140.
246. Boresch, S. and M. Karplus, *The Jacobian factor in free energy simulations.* J. Chem. Phys, 1996. **105**(12), 5145-5154.
247. Straatsma, T.P., M. Zacharias, and J.A. Mccammon, *Holonomic constraint contributions to free energy differences from thermodynamic integration molecular dynamics simulations.* Chem. Phys. Lett., 1992. **196**(3-4), 297-302.
248. Hornak, V., et al., *Comparison of multiple Amber force fields and development of improved protein backbone parameters.* Proteins: Struct. Funct. Bioinf., 2006. **65**(3), 712-25.
249. C.R. Calladine, H.R.D., B. Luisi, A.A. Travers, *Understanding DNA: the molecule and how it works.* Elsevier Academic Press, 2004.
250. Roberts, R.J. and X. Cheng, *Base flipping.* Annu. Rev. Biochem., 1998. **67**, 181-98.
251. Roberts, R.J., *On base flipping.* Cell, 1995. **82**(1), 9-12.
252. Gueron, M., M. Kochoyan, and J.L. Leroy, *A single mode of DNA base-pair opening drives imino proton exchange.* Nature, 1987. **328**(6125), 89-92.
253. Wu, J.C. and D.V. Santi, *Kinetic and catalytic mechanism of HhaI methyltransferase.* J. Biol. Chem., 1987. **262**(10), 4778-86.
254. Klimasauskas, S., et al., *Dynamic modes of the flipped-out cytosine during HhaI methyltransferase-DNA interactions in solution.* Embo Journal, 1998. **17**(1), 317-24.
255. Klimasauskas, S., et al., *HhaI methyltransferase flips its target base out of the DNA helix.* Cell, 1994. **76**(2), 357-69.
256. Reinisch, K.M., et al., *The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing.* Cell, 1995. **82**(1), 143-53.
257. Slupphaug, G., et al., *A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA.* Nature, 1996. **384**(6604), 87-92.
258. Chen, Y.Z., V. Mohan, and R.H. Griffey, *Spontaneous base flipping in DNA and its possible role in methyltransferase binding.* Phys. Rev. E Stat. Phys. Plasmas Fluids Relat Interdiscip Topics, 2000. **62**(1 Pt B), 1133-7.

259. Hagan, M.F., et al., *Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA.* Proc. Natl. Acad. Sci. USA, 2003. **100**(24), 13922-13927.
260. Klimasauskas, S., et al., *Hhal Methyltransferase Flips Its Target Base out of the DNA Helix.* Cell, 1994. **76**(2), 357-369.
261. Cheng, X.D. and R.J. Roberts, *AdoMet-dependent methylation, DNA methyltransferases and base flipping.* Nucleic Acids Res, 2001. **29**(18), 3784-3795.
262. Huang, N., N.K. Banavali, and A.D. MacKerell, *Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase.* Proc. Natl. Acad. Sci. USA, 2003. **100**(1), 68-73.
263. Yakovchuk, P., E. Protozanova, and M.D. Frank-Kamenetskii, *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nucleic Acids Res, 2006. **34**(2), 564-74.
264. Varnai, P. and K. Zakrzewska, *DNA and its counterions: a molecular dynamics study.* Nucleic Acids Res, 2004. **32**(14), 4269-80.
265. Sitkoff, D., K.A. Sharp, and B. Honig, *Accurate calculation of hydration free energies using macroscopic solvent models.* J. Phys. Chem., 1994. **98**, 1978-1988.
266. Cornilescu, G., et al., *Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase.* J. Am. Chem. Soc., 1998. **120**(27), 6836-6837.
267. Lavery, R. and H. Sklenar, *The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids.* J. Biomol. Struct. Dyn., 1988. **6**(1), 63-91.
268. Lu, X.J. and W.K. Olson, *3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.* Nucleic Acids Res, 2003. **31**(17), 5108-21.
269. Olson, W.K., et al., *A standard reference frame for the description of nucleic acid base-pair geometry.* J. Mol. Biol., 2001. **313**(1), 229-237.
270. Warshel, A., *Computer Modeling of Chemical reactions in enzymes and solutions.* John wiley & Sons, 1991.
271. Russu, I.M., *Probing site-specific energetics in proteins and nucleic acids by hydrogen exchange and nuclear magnetic resonance spectroscopy.* Methods Enzymol., 2004. **379**, 152-75.
272. Giudice, E., P. Varnai, and R. Lavery, *Energetic and conformational aspects of A : T base-pair opening within the DNA double helix.* ChemPhysChem, 2001. **2**(11), 673-677.
273. Giudice, E., P. Varnai, and R. Lavery, *Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations.* Nucleic Acids Res, 2003. **31**(5), 1434-1443.
274. Varnai, P., M. Canalia, and J.L. Leroy, *Opening mechanism of G center dot T/U pairs in DNA and RNA duplexes: A combined study of imino proton exchange and molecular dynamics simulation.* J. Am. Chem. Soc., 2004. **126**(44), 14659-14667.
275. Spies, M.A. and R.L. Schowen, *The trapping of a spontaneously "flipped-out" base from double helical nucleic acids by host-guest complexation with beta-cyclodextrin: The intrinsic base-flipping rate constant for DNA and RNA.* J. Am. Chem. Soc., 2002. **124**(47), 14049-14053.
276. Daujotyte, D., et al., *Hhal DNA methyltransferase uses the protruding Gln237 for active flipping of its target cytosine.* Structure, 2004. **12**(6), 1047-1055.
277. Chen, Y.Z., V. Mohan, and R.H. Griffey, *The opening of a single base without perturbations of neighboring nucleotides: A study on crystal B-DNA duplex d(CGCGAATTCGCG)(2).* J. Biomol. Struct. Dyn., 1998. **15**(4), 765-777.
278. Keepers, J.W., et al., *Molecular Mechanical Studies of DNA Flexibility - Coupled Backbone Torsion Angles and Base-Pair Openings.* Proc. Natl. Acad. Sci. USA, 1982. **79**(18), 5537-5541.
279. Ramstein, J. and R. Lavery, *Energetic Coupling between DNA Bending and Base Pair Opening.* Proc. Natl. Acad. Sci. USA, 1988. **85**(19), 7231-7235.
280. Briki, F., et al., *Evidence for the Stochastic Nature of Base Pair Opening in DNA - a Brownian Dynamics Simulation.* J. Am. Chem. Soc., 1991. **113**(7), 2490-2493.
281. P. Varnai, L.R., *Modelling DNA deformations, in Computational studies of DNA and RNA.* eds J. Sponder, and F. Lankas, Springer-Verlag, Berlin, 2006.
282. Bernet, J., K. Zakrzewska, and R. Lavery, *Modelling base pair opening: the role of helical twist.* Journal of Molecular Structure-Theochem, 1997. **398**, 473-482.
283. Huang, N. and A.D. MacKerell, Jr., *Atomistic view of base flipping in DNA.* Philos. Transact. A Math. Phys. Eng. Sci., 2004. **362**(1820), 1439-60.
284. Guerriertakada, C., et al., *The Rna Moiety of Ribonuclease-P Is the Catalytic Subunit of the Enzyme.* Cell, 1983. **35**(3), 849-857.
285. Kruger, K., et al., *Self-Splicing Rna - Auto-Excision and Auto-Cyclization of the Ribosomal-Rna Intervening Sequence of Tetrahymena.* Cell, 1982. **31**(1), 147-157.

286. North, G., *Nobel prizes: chemistry. RNA's catalytic role.* Nature, 1989. **341**(6243), 556.
287. Lilley, D.M.J., *The origins of RNA catalysis in ribozymes.* Trends in Biochemical Sciences, 2003. **28**(9), 495-501.
288. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution.* Science, 2000. **289**(5481), 905-920.
289. Winkler, W.C., S. Cohen-Chalamish, and R.R. Breaker, *An mRNA structure that controls gene expression by binding FMN.* Proc. Natl. Acad. Sci. USA, 2002. **99**(25), 15908-15913.
290. Nahvi, A., et al., *Genetic control by a metabolite binding mRNA.* Chemistry & Biology, 2002. **9**(9), 1043-1049.
291. Gollnick, P. and P. Babitzke, *Transcription attenuation.* Biochim. Biophys. Acta., 2002. **1577**(2), 240-50.
292. Henkin, T.M. and C. Yanofsky, *Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions.* Bioessays, 2002. **24**(8), 700-7.
293. Winkler, W., A. Nahvi, and R.R. Breaker, *Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression.* Nature, 2002. **419**(6910), 952-956.
294. McDaniel, B.A.M., et al., *Transcription termination control of the S box system: Direct measurement of S-adenosylmethionine by the leader RNA.* Proc. Natl. Acad. Sci. USA, 2003. **100**(6), 3083-3088.
295. Mandal, M., et al., *Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria.* Cell, 2003. **113**(5), 577-586.
296. Tucker, B.J. and R.R. Breaker, *Riboswitches as versatile gene control elements.* Curr. Opin. Struct. Biol., 2005. **15**(3), 342-8.
297. Nudler, E., *Flipping riboswitches.* Cell, 2006. **126**(1), 19-22.
298. Serganov, A., et al., *Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs.* Chem. Biol., 2004. **11**(12), 1729-41.
299. Montange, R.K. and R.T. Batey, *Structure of the S-adenosylmethionine riboswitch regulatory mRNA element.* Nature, 2006. **441**(7097), 1172-5.
300. Serganov, A., et al., *Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch.* Nature, 2006. **441**(7097), 1167-71.
301. Gilbert, S.D., et al., *Mutational analysis of the purine riboswitch aptamer domain.* Biochemistry, 2007. **46**(46), 13297-309.
302. Barrick, J.E., et al., *New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control.* Proc. Natl. Acad. Sci. USA, 2004. **101**(17), 6421-6426.
303. Mironov, A.S., et al., *Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria.* Cell, 2002. **111**(5), 747-56.
304. Hobartner, C., et al., *RNA two-state conformation equilibria and the effect of nucleobase methylation.* Angewandte Chemie-International Edition, 2002. **41**(4), 605-609.
305. Wenter, P., et al., *Kinetics of photoinduced RNA refolding by real-time NMR spectroscopy.* Angewandte Chemie-International Edition, 2005. **44**(17), 2600-2603.
306. Uhlenbeck, O.C., *Tetraloops and RNA folding.* Nature, 1990. **346**(6285), 613-4.
307. Chauhan, S. and S.A. Woodson, *Tertiary interactions determine the accuracy of RNA folding.* J. Am. Chem. Soc. , 2008. **130**(4), 1296-303.
308. Sattin, B.D., et al., *Direct measurement of tertiary contact cooperativity in RNA folding.* J. Am. Chem. Soc., 2008. **130**(19), 6085-7.
309. Wu, H., et al., *A novel family of RNA tetraloop structure forms the recognition site for Saccharomyces cerevisiae RNase III.* Embo Journal, 2001. **20**(24), 7240-9.
310. Brion, P. and E. Westhof, *Hierarchy and dynamics of RNA folding.* Annu. Rev. Biophys. Biomol. Struct., 1997. **26**, 113-37.
311. Freier, S.M., et al., *Improved Free-Energy Parameters for Predictions of Rna Duplex Stability.* Proc. Natl. Acad. Sci. USA, 1986. **83**(24), 9373-9377.
312. Turner, D.H., et al., *Improved Parameters for Prediction of Rna Structure.* Cold Spring Harbor Symposia on Quantitative Biology, 1987. **52**, 123-133.
313. Xia, T.B., et al., *Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs.* Biochemistry, 1998. **37**(42), 14719-14735.
314. Borer, P.N., et al., *Stability of Ribonucleic-Acid Double-Stranded Helices.* J. Mol. Biol., 1974. **86**(4), 843-853.
315. Tinoco, I., et al., *Improved Estimation of Secondary Structure in Ribonucleic-Acids.* Nature-New Biology, 1973. **246**(150), 40-41.

316. Zuker, M. and P. Stiegler, *Optimal Computer Folding of Large Rna Sequences Using Thermodynamics and Auxiliary Information.* Nucleic Acids Res, 1981. **9**(1), 133-148.
317. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13), 3406-3415.
318. Gruber, A.R., et al., *The Vienna RNA Websuite.* Nucleic Acids Res, 2008. **36**(Web server issue), W70-W74.
319. Ding, Y., C.Y. Chan, and C.E. Lawrence, *Sfold web server for statistical folding and rational design of nucleic acids.* Nucleic Acids Res, 2004. **32**(Web server issue), W135-W141.
320. Steffen, P., et al., *RNAshapes: an integrated RNA analysis package based on abstract shapes.* Bioinformatics, 2006. **22**(4), 500-503.
321. Xayaphoummine, A., T. Bucher, and H. Isambert, *Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots.* Nucleic Acids Res, 2005. **33**, W605-W610.
322. Isambert, H. and E.D. Siggia, *Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme.* Proc. Natl. Acad. Sci. USA, 2000. **97**(12), 6515-6520.
323. Chen, X.Y., et al., *A characteristic bent conformation of RNA pseudoknots promotes -1 frameshifting during translation of retroviral RNA.* J. Mol. Biol., 1996. **260**(4), 479-483.
324. Namy, O., et al., *A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting.* Nature, 2006. **441**(7090), 244-247.
325. Hainzl, T., S.H. Huang, and A.E. Sauer-Eriksson, *Structure of the SRP19-RNA complex and implications for signal recognition particle assembly.* Nature, 2002. **417**(6890), 767-771.
326. Ennifar, E., et al., *The crystal structure of UUCG tetraloop.* J. Mol. Biol., 2000. **304**(1), 35-42.
327. Ryckaert, J.P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes.* J. Comput. Phys., 1977(23), 327-341.
328. Allain, F.H. and G. Varani, *Structure of the P1 helix from group I self-splicing introns.* J. Mol. Biol. , 1995. **250**(3), 333-53.
329. Mlynsky V, et al., *Extensive Molecular Dynamics Simulations Showing That Canonical G8 and Protonated A38H+ Forms Are Most Consistent with Crystal Structures of Hairpin Ribozyme.* J. Phys. Chem. B, 2010. **114**(19), 6642-6652.
330. Eyring, H., *The activated complex in chemical reactios.* J. Chem. Phys., 1935. **3**(2), 107-115.
331. Antao, V.P. and I. Tinoco, Jr., *Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops.* Nucleic Acids Res, 1992. **20**(4), 819-24.
332. Schmitz, U., et al., *Structure of the phylogenetically most conserved domain of SRP RNA.* RNA, 1999. **5**(11), 1419-29.
333. Jucker, F.M., et al., *A network of heterogeneous hydrogen bonds in GNRA tetraloops.* J. Mol. Biol. , 1996. **264**(5), 968-80.
334. Batey, R.T., et al., *Crystal structure of the ribonucleoprotein core of the signal recognition particle.* Science, 2000. **287**(5456), 1232-9.
335. Macke, T.J., et al., *Amber Tools.* 2008.
336. Banavali, N.K. and B. Roux, *Free energy landscape of A-DNA to B-DNA conversion in aqueous solution.* J. Am. Chem. Soc. , 2005. **127**(18), 6866-76.
337. Khavrutskii, I.V., J. Dzubiella, and J.A. McCammon, *Computing accurate potentials of mean force in electrolyte solutions with the generalized gradient-augmented harmonic Fourier beads method.* J. Chem. Phys., 2008. **128**(4), 044106.
338. Du R, et al., *On the transition coordinate for protein folding.* . J. Chem. Phys. , 1998. **108**(1), 334-350.
339. Hubner, I.A., M. Oliveberg, and E.I. Shakhnovich, *Simulation, experiment, and evolution: understanding nucleation in protein S6 folding.* Proc. Natl. Acad. Sci. USA, 2004. **101**(22), 8354-9.
340. Morris, G.A. and R. Freeman, *Selective excitation in Fourier transform nuclear magnetic resonance.* J. Magn. Res, 1978. **29**(3), 433-462.
341. Snoussi, K. and J.L. Leroy, *Imino proton exchange and base-pair kinetics in RNA duplexes.* Biochemistry, 2001. **40**(30), 8898-904.
342. Ernst, R.R., G. Bodenhausen, and A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions.* International Series of Monographs on Chemistry, Oxford: University Press, 1987. **14**.
343. Sorin, E.J., et al., *Insights into nucleic acid conformational dynamics from massively parallel stochastic simulations.* Biophys. J., 2003. **85**(2), 790-803.

344. Bowman, G.R., et al., *Structural insight into RNA hairpin folding intermediates.* J. Am. Chem. Soc., 2008. **130**(30), 9676-9768.
345. Zhuang, Z.Y., L. Jaeger, and J.E. Shea, *Probing the structural hierarchy and energy landscape of an RNA T-loop hairpin.* Nucleic Acids Res, 2007. **35**(20), 6995-7002.
346. Deng, N.J. and P. Cieplak, *Free Energy Profile of RNA Hairpins: A Molecular Dynamics Simulation Study.* Biophys. J., 2010. **98**(4), 627-636.
347. Weeks, J.D., D. Chandler, and H. Andersen, *Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids* J. Chem. Phys., 1971. **54**, 5237-5247.
348. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600), 1912-34.
349. Huse, M. and J. Kuriyan, *The conformational plasticity of protein kinases.* Cell, 2002. **109**(3), 275-282.
350. Morgan, D.O. and H.L. Debondt, *Protein-Kinase Regulation - Insights from Crystal-Structure Analysis.* Curr. Opin. Cell Biol., 1994. **6**(2), 239-246.
351. Levinson, N.M., et al., *A Src-like inactive conformation in the Abl tyrosine kinase domain.* Plos Biology, 2006. **4**(5), 753-767.
352. Rowley, J.D., *Biological Implications of Consistent Chromosome Rearrangements in Leukemia and Lymphoma.* Cancer Research, 1984. **44**(8), 3159-3168.
353. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* J. Comput. Chem., 2005. **26**(16), 1781-802.